THE UNIVERSITY
of EDINBURGH

Characterization of genome-wide deviations from Mendelian inheritance in bivalve species

Carolina Soledad Peñaloza Navarro

Doctor of Philosophy

University of Edinburgh

2017

Abstract

Marine bivalves are a group of species composed of clams, mussels and oysters. Bivalves are keystone species in coastal ecosystems and represent an increasingly important segment of the global aquaculture industry. Domestication of shellfish species is in the early stages, with few organized breeding programmes and a heavy reliance on wild seed. Consequently, the development and use of genomic markers may significantly assist shellfish aquaculture breeding and production. However, molecular genetic markers typically exhibit unusual patterns of segregation in bivalve species, which result in deviations from Mendelian expectations, and could potentially limit their use in parental assignment, mapping of quantitative trait loci and genomic prediction. Previous studies have suggested that segregation distortions originate at the larval stage, as a result of the linkage of markers to deleterious mutations. This high genetic load has been associated with the high fecundity of bivalve species. However, no direct evidence of a high incidence of *de novo* mutations has been provided. The aim of this thesis is to gain further insight into segregation distortions in bivalve species by studying the phenomenon at a genome-wide scale, using modern high-throughput sequencing technology. The studies presented in this thesis derive from experiments involving genotyping of parents and offspring from pair-crosses of three different bivalve species (the Pacific oyster *Crassostrea gigas*, the Blue mussel *Mytilus edulis*, and the Greenshell™ mussel *Perna canaliculus*) using high throughput sequencing and SNP arrays. The parent and offspring genotype data were used to characterize patterns of segregation distortion at a genome-wide level, followed by exploratory analyses to test hypotheses related to possible causes of this distortion. Three main findings resulted from the genome-wide analysis of segregation patterns. First, by using Restriction site Associated DNA sequencing (RAD-Seq) we observe that technical artefacts are more widespread than previously considered, contributing to apparent distortions via unreliable genotype calls. By analysing read depth data from RAD-Seq, we suggest that apparent homozygous genotype calls may actually be hemizygous, suggesting a very high frequency of null alleles which contribute to distorted segregation patterns. Bioinformatic pipelines to improve RAD-Seq locus assembly and marker genotyping for bivalve species are presented. Second, by using a high-density SNP array and RAD-Seq in pair crosses of Pacific oyster and aligning to the reference genome assembly, we find that segregation

distortions cover extensive regions of the genome, and that certain genomic regions are consistently distorted in different families. Finally, following previous suggestions that the reproductive strategies of bivalve species may favour a high mutation rate, we provide preliminary evidence of a high incidence of *de novo* mutations that appear spontaneously (i) during male and female gamete formation and (ii) post-zygotically, during larval development. This putative high *de novo* mutation rate is likely to also contribute to deviations from Mendelian inheritance patterns in these species. New genomic technologies have allowed us to gain substantial insight into the intriguing yet poorly understood phenomena related to inheritance in bivalve species. The results have both fundamental and practical implications for genetic analysis interpretation and selective breeding for aquaculture in this large and highly diverse group of species.

Lay Summary

Marine bivalves are a group of species composed of clams, mussels and oysters. Bivalves (i.e., animals with two valves) are keystone species in coastal ecosystems and represent an increasingly important segment of the global aquaculture industry. Domestication of bivalve shellfish species is in the early stages, with few organized breeding programmes and a heavy reliance on wild juveniles to sustain the production. The quality and quantity of these juveniles varies inter-annually. Therefore, efforts are being made to start producing these juveniles in hatcheries (nurseries) through artificial fertilisation. If we have control of reproduction, then we will have an opportunity to genetically improve farmed bivalves through selective breeding. However, there is a potential issue with selective breeding in bivalves. The basis of selective breeding is the stable transmission of genetic information from parents to offspring. Bivalves appear to deviate from this expectation and we do not understand why. The general aim of this thesis is to explore the reasons for the unusual inheritance patterns of bivalves. For my research, I created mussel and oyster families and used high-throughput sequencing technologies to study genetic transmission (from parents to offspring) at a high genomic resolution. The results of this work have both fundamental and practical implications for genetic analysis interpretation and selective breeding for aquaculture in this large and highly diverse group of species.

ACKNOWLEDGEMENTS

Firstly, I would like to thank the Chilean people for giving me the scholarship that allowed me to develop this thesis. Secondly, I would like to thank my supervisors Steve Bishop, Chris Haley and Ross Houston for their guidance. Particularly I would like to thank Ross Houston because he gave me invaluable scientific freedom and support. I would also like to thank Andy Law for being a great thesis Committee Chair. I am grateful to my collaborators at Cefas (UK) Tim Bean, and Cawthron (NZ), Serean Adams, Jane Symonds and Bridgett Finnie, who taught me how to raise mussels and oysters; and to John Taggart (University of Stirling, UK) for giving me the tilapia samples that I used as controls. Also, I would like to acknowledge the people at the BRF, who helped me in my initial attempts of isolating single gametes. Thank you to all the people in the Roslin Institute, particularly to the Houston group. The biggest thank is for Marcelo who shared this path with me, thank you for your infinite patience. Finally, I would like to acknowledge the people who develop free software and the bioinformaticians that share their knowledge in places such as Stack Overflow; thank you for being my teachers.

Abbreviations

bp	base pairs
DNA	Deoxyribonucleic acid
^{o}C	Celcius
CTAB	Cetyl trimethylammonium bromide
EB	elution buffer
EDTA	Ethylenediaminetetraacetic acid
g	Relative centrifuge force
Mb	Mega base pair
min	minutes
ml	millilitre
mm	millimetre
mM	millimolar
NaCl	Sodium chloride
nm	nanomolar
ph	potential of hydrogen
ppt	parts per thousand
RAD-Seq	Restriction site Associated DNA Sequencing
rpm	revolutions per minute

RNAse	ribonuclease
s	seconds
SW	sea water
Tris-HCl	Tris(hydroxymethyl)aminomethane hydrochloride
uM	micromolar
vcf	variant calling format

Table of Contents

CHAPTER 2

General Material and Methods

Evaluation of de novo assembly and variant calling methods for bivalve RAD-Seq data

CHAPTER 4

Characterization of genome-wide patterns of segregation distortion in Pacific oyster full-sibling families

CHAPTER 7

General discussion

CHAPTER 1

General Introduction

1.1 Introduction

Mendel inferred the principles that govern inheritance and explain how traits in offspring can be predicted from traits in their parents, the Mendelian laws of inheritance (reviewed in Singh (2015)). Soon after Mendel's scientific breakthrough was recognized, the first example of non-Mendelian inheritance was published (reviewed in Sakamoto et al. (2008)). Although most traits in animals and plants follow Mendel's laws of heredity, some traits and species show more complex patterns of inheritance. Particular examples of non-Mendelian inheritance include the t-alleles of mice and the SD-chromosomes of Drosophila (Taylor and Ingvarsson, 2003). Remarkable cases of apparent non-Mendelian inheritance are observed in marine bivalves, with frequent reports of unexpected genetic marker segregation patterns (Plough, 2016a) and inconsistencies between parent and offspring genotypes (Peñaloza, 2013). Despite decades of research into bivalve genetics, our understanding of why these unusual segregation patterns emerge is still very limited.

Bivalves are a highly diverse class of molluscs comprised of mussels, oyster, clams, and scallops. They exhibit a high diversity of reproduction strategies, which are influenced by both genetics and environmental factors (Breton et al., 2017). Bivalve species are generally gonochoric (i.e., have both sexes), although sequential hermaphroditism also occurs (Heller, 1993). Most bivalves reproduce externally by synchronously releasing gametes into the water during massive spawning events. After fertilization, larvae drift in the water column for 3-5 weeks. Subsequently, they settle onto a suitable substrate and remain virtually immobile for the rest of their life, filtering food particles from the water (Gosling, 1992). The biphasic life-history of these organisms is suggested to have profound consequences for their evolution, as selective pressures on the different stages are potentially decoupled, therefore limiting their evolution. This implies that, to evolve, traits and their underlying genotypes must be simultaneously adapted to the radically different ecological contexts that organisms experience during their ontogeny (Raff, 1987, Schluter D. et al., 1991, Marshall and Morgan, 2011). In spite of these evolutionary constraints, bivalves have successfully colonized a wide range of environments. They inhabit the temperate and boreal waters of both hemispheres (Gosling, 1992), and can be found in tidal and sub-tidal waters, from rocky coasts to deep-sea hydrothermal vents at >3,000 m depth (Duperron et al., 2006). Moreover, they can survive in heavily contaminated areas (Kim et al., 2002), and spread into new environments as invasive species (Marescaux and

Van Doninck, 2013). As part of their mechanisms to cope with harsh marine environments, bivalves have developed a sophisticated immune system that is based on an expanded repertoire of key genes (Guo et al., 2015). The expansion of specific gene families has been suggested to be the result of adaptation to fluctuations in environmental stressors. For instance, the Pacific oyster *Crassostrea gigas* genome contains 88 heat shock protein 70 Hsp70 (group of proteins associated with the stress response) compared to ~17 in humans (Zhang et al., 2012). The phylogenetic analysis of the Hsp70 proteins from the Pacific oyster and the pearl oyster *Pinctada fucata* indicates that the extensive expansion of genes occurred prior to the divergence of the two species, but also arose independently in each lineage (Takeuchi et al., 2016), suggesting that tandem duplication events are common in bivalves and have facilitated their adaptation to stressful environments.

Bivalves also fulfil critical ecological and economic roles in coastal systems (Dumbauld et al., 2009). As filter-feeders, the functional role of bivalves is to influence benthic-pelagic coupling and nutrient cycling (Newell, 2004). As aquaculture resources, they represent 15% of the world production (FAO, 2014), with a high potential for expansion (Gentry et al., 2017). Due to their importance for ecosystems and human consumption, in addition to their use as model organisms to study the evolution of innate immunity and stress responses (Guo et al., 2015, Liu et al., 2016, Zhang et al., 2012), a significant amount of genomic and transcriptomic data is being generated. However, the development of tools needs accompanying progress in a fundamental aspect of bivalve biology that remains unresolved, namely a fuller understanding of their mechanisms of genetic transmission and inheritance.

1.1.1 Bivalve Inheritance

In the cell of eukaryotic organisms, genetic information is present in two distinct forms – as nuclear or mitochondrial DNA. Nuclear and mitochondrial genomes differ in several ways (Taylor and Turnbull, 2005); however, the most distinctive feature is their mode of inheritance. Nuclear DNA is transmitted both from the paternal and maternal lineage, and is expected to adhere to Mendelian laws; whereas mitochondrial DNA is strictly inherited from the maternal line in a relatively clonal fashion (Birky et al., 1978), although leakage has been described (Dokianakis and Ladoukakis, 2014). Bivalves are remarkable in the sense

that they do not appear to conform – either at the mitochondrial or nuclear level – to these nearly universal patterns of inheritance.

1.1.2 Mitochondrial Inheritance

Mitochondria are responsible for producing most of the energy (>90%) required to run the cell. This function is achieved by converting lipids and carbohydrates into a high-energy molecule called adenosine triphosphate (ATP), through the process of oxidative phosphorylation (OXPHOS). The number of mitochondria within a cell varies according to cell type, reflecting the direct relationship between cell function and energy (ATP) demand. Consequently, cells such as those in the liver and muscle tend to have larger numbers of mitochondria (e.g., 1000-2000 per liver cell in humans), whereas mammalian erythrocytes completely lack mitochondria. Mitochondria are unique compared to other animal eukaryotic organelles as they contain their own DNA. Mitochondrial DNA (mtDNA) is circular, double-stranded and haploid, and contains genes that encode proteins involved in the OXPHOS process. A key characteristic of mtDNA is its higher mutation rate compared to their nuclear counterpart. The mitochondrial genome is more prone to nucleotide change mainly due to (i) the highly oxidative environment produced during ATP generation, (ii) an error-prone replication system (in higher metazoans), and (iii) the limited repair function of mitochondria (reviewed in Singh et al. (2012)). As a result, mutations occur at a rate approximately ten times higher in animal mitochondrial genomes than in equivalent nuclear genomes (Brown et al., 1979).

In most animal species mtDNA is maternally transmitted, where the mtDNA transferred to the next generation is the population of molecules present in the oocyte just prior to fertilization. Maternal inheritance is achieved by preventing paternal transmission of mtDNA through different species-specific mechanisms. For instance, in some tunicates (sea squirts) the mtDNA of the spermatozoa does not enter the oocyte (Rokas et al., 2003). In comparison, during human fertilization the paternal mtDNA penetrates the egg but within a few hours is selectively tagged with a proteolytic marker and is destroyed by the ubiquitin-proteasome system (Chan and Schon, 2012). In bivalve species of the families Mytilidae and Unionidae (sea and fresh water mussels, respectively) however, empirical evidence indicates that mtDNA is inherited both from the maternal and paternal line, although it

appears not to be a universal mechanism, as exceptions have been described (Obata et al., 2008).

This system of mtDNA inheritance from both parents is known as Double Uniparental Inheritance (DUI) (Skibinski et al., 1994, Zouros et al., 1992, Zouros et al., 1994), and represents a complex exception to the central tenets of mtDNA inheritance. Under DUI two gender-associated mitochondrial genomes coexist, within an individual, in a theoretically stable state. One mitochondrial genome, known as maternal (symbolized as F), is transmitted from mothers to both female and male progeny. The other mitochondrial DNA is paternal (symbolized as M), and is transmitted from males solely to their male offspring. This leads to the female offspring being homoplasmic (within an individual all mtDNA is near identical) for the F genome and the male offspring being heteroplasmic (individuals have more than one type of mtDNA, in this particular case: F and M mtDNA molecules). Furthermore, in sexually mature males the distribution of F and M genomes is tissue specific. The F mtDNA is predominant in somatic tissues, whereas the gonadal tissue is dominated by the M genome, although lower amounts of F mtDNA have also been observed in the gonad of male Blue mussels (*Mytilus edulis*) (Garrido-Ramos et al., 1998). Due to the fact that the M mtDNA establishes itself in the gonad of males, the transmission of the M genome to the next generation is ensured.

A possible mechanism to explain why M mtDNA is abundant in the male gonad and how it colonizes this tissue was proposed by Cao et al. (2004). The sex-ratio determination in mussels is, at least partially, under the control of the nuclear DNA of the female parent (Saavedra 1997); some dams produce mainly female offspring, whereas others produce a high frequency of male offspring (Kenchington et al., 2002). Cao and co-workers performed mussel crosses using females that produce offspring biased towards either sex and then followed the fate of sperm mitochondria in the developing embryo. They noticed that in embryos from females that produce mainly daughters, the M mtDNA (from the male donor) dispersed randomly among cells of the zygote, and became diluted in the F mtDNA pool of the embryo. In contrast, in embryos from male-biased mothers, the paternal mitochondria formed an aggregate that was maintained as a cluster until the termination of the experiment at 72 hours post fertilization. Based on this behavior, the researchers proposed that M mtDNA (once inside a male-to-be embryo) follow a pre-defined path towards the cell aggregates that will originate germline primordial cells (i.e., the future gonad).

Interestingly, DUI is not a robust inheritance mechanism, and many disruptions have been described. For instance, M-type mtDNA has been detected in the adult tissue of several females (Garrido-Ramos et al., 1998, Kyriakou et al., 2010). Moreover, in some occasions, the maternally transmitted F mtDNA can invade the paternal route of transmission. This event is referred to as the *masculinization* of the maternal mtDNA, which leads to its transmission to the next-generation as a new M-type (Breton et al., 2006, Burzynski et al., 2006, Hoeh et al., 1997). Other examples of disrupted DUI include homoplasmic males for the F genome (Fisher and Skibinski, 1990) and the presence of M and F mtDNA in unfertilized eggs (Obata et al., 2008). It can be concluded from the present knowledge on DUI that this is an extremely dynamic mechanism of inheritance, and many knowledge gaps remain.

1.1.3 Nuclear Inheritance in Bivalves

Most bivalve species studied to date are diploids (Thiriot-Quievreux, 2002). Nuclear (chromosomal) inheritance in sexual diploid organisms is expected to follow Mendelian segregation, resulting in each homologous chromosome being present in 50% of the gametes on average. Following this principle, genotypic ratios in the offspring can be predicted from the genotypes of their parents. The deviation from Mendelian expectation is known as segregation distortion (SD) and has been observed in a wide range of organisms, but is particularly frequently in plant and fungi species (Taylor and Ingvarsson, 2003).

SD is a hallmark of bivalve genetics, and has been identified in oyster (Zouros et al., 1983, Saavedra et al., 1993); scallop (Beaumont and Beveridge, 1984); mussel (Koehn and Gaffney, 1984, Diehl and Koehn, 1985, Beaumont, 1991); and clam species (David et al., 1997, Borsa et al., 1991). The widespread nature of SD was an unexpected finding, as it is inconsistent with bivalves' life history traits. Most marine bivalves show high fecundities, large population sizes, and extensive larval dispersal (Barnes and Ruppert, 1994), thus genotypic proportions within populations should conform to – rather than deviate from – Hardy-Weinberg expectation of ratios (although see Hedgecock and Pudovkin (2011)). Insights into the origin of SD in bivalves were enabled when marker segregation ratios were evaluated in progenies of experimental crosses. Evidence from family-based experiments

indicates that the tendency for SD differs between individuals from inbred crosses compared to the progeny from wild-parents, suggesting that the underlying mechanism requires a more thorough exploration.

Substantial SD is present in inbred crosses (~15-50% of markers assayed; Plough (2016a)), with a general tendency to observe a deficiency of homozygote genotypes in the offspring with respect to the expected Mendelian ratios (McGoldrick and Hedgecock, 1997). Given that distortions were exacerbated in inbred families (McGoldrick and Hedgecock, 1997), increased with the age of the animal (Zouros et al., 1983, Zouros et al., 1980), and showed no consistent pattern across families or markers (McGoldrick and Hedgecock, 1997, Mallet et al., 1985), the deficiency of homozygote genotypes was suggested to be caused by selection against deleterious recessive mutations at linked loci. By default, this would imply that fitness-related genes are evenly scattered throughout the genome. A further landmark study performed by Bierne et al. (1998) provided the empirical evidence that showed that the origin of SD was due to genotype-dependant mortalities. Two crosses between full-siblings were created, and by genotyping offspring with four microsatellite markers at different ages (including larvae and juveniles), the timing of distortions was established. When markers were tested for conformity to Mendelian inheritance, genotype ratios conformed to Mendelian expectations at early larval stages, while showing distortions towards a deficiency of homozygote genotypes (or conversely, excess of heterozygotes) in later sampled time points. The results suggested that SD was indeed caused by a high genetic load that became unmasked by the partially inbred state of the animals used in the study. Genetic load is the presence of unfit genotypes in a population compared to a theoretical population optimum (Whitlock and Davis, 2001). The mechanism by which genetic load may cause a reduction in the mean fitness of a population may be, for example, recurrent deleterious mutations. Plough (2016a) has proposed that bivalves may exhibit high mutation rates based, among other evidence, on their remarkably high ratio of non-synonymous to synonymous substitutions in oysters (e.g., dn/ds ratios =0.38 averaged across 37 loci; Harrang et al. (2013)). The dn/ds ratio is used to measure the strength and mode of natural selection pressures on protein-coding regions (Jeffares et al., 2015). Because amino acid changes will tend to have a detrimental effect on fitness, they will be removed by purifying selection. Therefore, non-synonymous mutations are expected to occur at a low frequency relative to synonymous mutations. Since the observation of a high dn/ds ratio in oysters is coupled with an excess of non-synonymous mutations segregating

at low frequencies (Harrang et al., 2007), a large fraction of weakly selected mutations was suggested to be segregating in wild populations, thus supporting a high genetic load (Plough, 2016a).

In the case of bivalve families created from unrelated parents (i.e., wild-caught parents) the patterns of SD tend to be different from those described for inbred families. Among the few studies available, all indicate that SD is mostly caused by a deficiency of the heterozygote genotype ratio with respect to Mendelian expectations (instead of the homozygote deficiency observed in inbred families). It is important to note that the majority of early studies in natural populations of bivalves agree with these findings, as they revealed that deviations from Hardy Weinberg equilibrium were mainly due to heterozygote deficiencies (Beaumont, 1991, Foltz, 1986, Raymond et al., 1997, Borsa et al., 1991, Toro and Vergara, 1995, Gosling and Wilkins, 1985, Lassen and Turano, 1978). Mallet et al. (1985) analyzed 49 blue mussel full-sibling families with six allozyme markers. An overall heterozygote deficiency was detected and was attributed to differences in larval viability or gametic selection (non-random fertilization). In a study performed by Plough et al. (2016), four pair-crosses of wild Pacific oysters were created, and offspring aged 1.4 and 2.3 years were genotyped with microsatellites and SNP markers. Results from the segregation analysis indicated that 97% of the microsatellite markers that showed distortions were because of a deficiency of heterozygote genotypes in the offspring. The use of highly polymorphic microsatellites markers allowed the researchers to estimate paternal effects associated with genotype deficiencies. By analysing crosses that allowed the authors to unambiguously associate alleles with linked deleterious mutations (e.g., in the case of the cross type: "ABxCD"→ expectation in offspring should follow a 1:1:1:1 ratio), they observed that 66% of the distorted markers were deficient in 3 or 4 genotype categories, suggesting that independent deleterious mutations were segregating from both parents under the assumption of larval selection being the cause of SD. Another interesting observation from this study was that to explain the substantial amounts of individuals “assumed” to carry distorted genotypes, ~99% of the progeny had to have died. This is in striking contrast with what is expected from an outbred family under the hypothesis that bivalves carry significant numbers of deleterious recessive alleles, as they would tend to be masked in a highly heterozygous genome. Moreover, previous studies on inbred Pacific oyster families show similar mortalities, 90-96% (Plough, 2012, Plough and Hedgecock, 2011). High mortalities in outbred Pacific oyster families was interpreted by the authors as the

consequence of partially dominant deleterious mutations. Overall, this illustrates that the genetic architecture of SD varies depending on the genetic background of individuals (inbreed vs. cross-bred), suggesting that a better understanding of the mechanisms underlying SD patterns in bivalves is required.

1.1.4 Cytogenetics

Data on standard karyotypes indicate that the majority of mollusc species studied to date are diploid, with the most frequent chromosome number being 2n = 38, indicating it may represent the ancient karyotype of bivalves (Thiriot-Quievreux, 2002). Nevertheless, most aquaculture bivalve species with karyological data show a tendency towards lower chromosome numbers (e.g., 2n=20 in the Pacific oyster and 2n=28 in *Mytilus* mussels; Thiriot-Quiéveux and Ayraud (1982)). The genome content of bivalve mollusks varies greatly across species. Comparative analysis indicates that the DNA content of the haploid nucleus (termed C-value) ranges from 0.9 pg in the Pacific oyster to 3.4 pg in the horse mussel (Thiriot-Quievreux, 2002). Notably, apart from this wide inter-specific variation, intra-specific variation in genome-size has also been reported, suggesting that at least part of the genome of bivalves is free to vary among individuals of the same species without major biological consequences (Rodriguez-Juiz et al., 1996, Martínez-Lage et al., 1997). Indeed, a striking feature of the bivalve genome is the frequent natural occurrence of aneuploidy (Longwell and Stiles, 1968, Thiriot-Quiévreux et al., 1992).

Aneuploidy refers to the loss or gain of one or more chromosomes with respect to the normal chromosome set (Dey, 2004). Although levels of chromosomal imbalance are generally associated with disease or lethality in higher animals (Korenberg et al., 1994, Hassold et al., 2007), bivalves can tolerate significant levels of unbalanced chromosome numbers. Abnormalities in chromosome number can result from errors in cell division such as homologs failing to cross over (MacLennan et al., 2015) or defects in the spindle check-point (Bharadwaj and Yu, 2004), and are expressed by cells exhibiting hypodiploidy or hyperdiploidy. Hypodiploidy is the condition in which fewer chromosomes are detected with respect to the baseline diploid set. Conversely, hyperdiploidy indicates an excess in the number of chromosomes. In a study performed in Pacific oyster individuals from a natural population, hypodiploidy was found in up to 33% of the cells, although most commonly was

found to affect 14% of cells (Zouros et al., 2009). Additionally, in this study, a negative correlation between somatic aneuploidy and body size was found. The authors hypothesised that (i) the high occurrence of aneuploid somatic cells and (ii) the negative correlation between the individual's degree of aneuploidy and growth rate was caused by a high mutation rate in the germline due to high gamete output (i.e., high fecundity). If these mutations affected mitosis-related genes, then spontaneous chromosomal loss would occur during development. This loss of DNA would lead to the unmasking of fitness-related mutations, consequently causing a detrimental effect on growth. Because they found no single individual with aneuploidy in all examined cells, they concluded that the mitosis-regulating mutations have a post-zygotic function and that 100% aneuploid zygotes are likely to be unviable. Another characteristic of cytogenetic abnormalities in bivalves is that they become exacerbated in contaminated areas. For instance, oysters were shown to increase their levels of aneuploidy in response to an herbicide of the triazine class (atrazine) (Bouilly et al., 2003). Aneuploidy in the group of juveniles exposed to the highest concentration of atrazine (465 nM) ranged from 24.3% to 25.3%, whereas in the control treatment ranged from 9.7% to 10.3%. In another study by Dixon (1982), aneuploidy was reported in ~8% of the embryos of laboratory-spawned mussels. However, this frequency rose to 26 % when the mussels were from a hydrocarbon-polluted area. Notably, a genetic basis for levels of aneuploidy has been suggested, as a significant difference in the degree of aneuploidy has been detected among families raised under the same environmental conditions (Leitão et al., 2001); the aneuploidy level within the family with the lowest mean weight was 36%, a value significantly higher compared to the level of 13% detected for the family with the highest mean weight.

1.1.5 Genomic resources

Genomic resources are valuable tools for ecological and evolutionary studies, and for the enhancement of traditional breeding programs through genomic selection (Meuwissen et al., 2001). The advances in next-generation sequencing (NGS) have contributed enormously to the development of genomic resources in bivalve species, providing an unprecedented opportunity to examine genomes at a high-resolution. Genomic resources in bivalve molluscs do not tend to be associated with the economic importance of the group of species. Despite the fact that clams lead aquaculture production (FAO, 2014), only recently

a high-density linkage map has become available for a clam species, the Manila *clam Ruditapes philippinarum* (Nie et al., 2017). Instead, oysters have the majority of developed resources, among which the Pacific oyster *Crassostrea gigas* has received most attention (Astorga, 2014). The genomic resources of oysters include BAC libraries (Cunningham et al., 2006), linkage maps for several species (Hubert and Hedgecock, 2004, Yu and Guo, 2003, Yaohua et al., 2009, Hedgecock et al., 2015, Li and Guo, 2004, Lallias et al., 2007), genome-wide SNP markers (Jones et al., 2013), and SNP arrays (Gutierrez et al. 2017, Qi et al. 2017). For mussels, on the other hand, a low-density AFLP linkage map is available for the Mediterranean mussel *M. galloprovincialis* (Lallias et al., 2007). Additionally, a microsatellite linkage map has also been produced for the triangle pearl mussel *Hyriopsis cumingii* (Bai et al., 2015). For scallops, fosmid libraries (Zhang et al., 2007) and linkage maps (Jiao et al., 2014, Yuan et al., 2010) are the main resources. Although not strictly a genomic resource, transcriptome data have been produced for several species, and were used to gain insight into bivalves: (i) tolerance to strong temporal changes in environmental conditions (typical of coastal environments), such as prolonged hypoxia (Sussarellu et al., 2010) and salinity stress (Zhao et al., 2012); (ii) growth heterosis (Hedgecock et al., 2007); (iii) shell formation (Yarra et al., 2016, Zhang et al., 2012, Clark et al., 2010); (iv) sex determination (Zhang et al., 2014); and (v) susceptibility to viral (Segarra et al., 2014) and bacterial diseases (de Lorgeril et al., 2011). Additionally, given the use of bivalve molluscs to monitor water-quality as sentinel species, the transcriptomics response to pollutants has also been evaluated (Dondero et al., 2011). Currently, transcriptome databases are publically available for the Mediterranean mussel (Venier et al., 2009), the Pacific oyster (Riviere et al., 2015), the Manila Clam (RuphiBase), the striped venus clams *Chamelea gallina* (Coppe et al., 2012), and the deep-sea hydrothermal vent mussel *Bathymodiolus azoricus* (Bettencourt et al., 2010).

A significant advance in the study of a species is to obtain a full representation of its genome, as it reveals important organizational, structural, and functional features of genes (Berg, 2006). Draft genomes have been published for the Pacific oyster *Crassostrea gigas* (Zhang et al., 2012), the Pearl oyster *Pinctada fucata* (Takeuchi et al., 2012, Takeuchi et al., 2016), the Yesso Scallop *Patinopecten yessoensis* (Wang et al., 2017), a deep-sea mussel *Bathymodiolus platifrons* (Sun et al., 2017), the Philippine horse mussel *Modiolus philippinarum* (Sun et al., 2017), and the Mediterranean mussel *Mytilus galloprovincialis* (Murgarella et al., 2016), although the latter was highly fragmented. In general, the

assembly of most bivalve genomes is a complex task, mainly because with high levels of heterozygosis it is often difficult to distinguish between polymorphisms at the same locus (in the case of divergent alleles) or a similar sequence at a different genomic location (in the case of paralogues). In particular, greater complications are experienced if a de Bruijn graph is used for assembling short sequencing reads, as highly heterozygous genomes produce a higher density of variant branching in the assembly graph (Simpson, 2014). A common strategy to circumvent the issue of genome heterozygosity is to reduce the genetic diversity of an organism through selective (in)breeding. This approach has proven useful in the assembly of bivalve genomes. For instance, Zhang et al. (2012) sequenced and assembled a Pacific oyster individual derived from four generations of full-sibling mating. Nevertheless, the reduction in heterozygosity was lower than the predicted from the mating scheme (expected polymorphism rate: 46%, observed: 73%), an observation that was interpreted as natural selection favoring heterozygotes. On the other hand, for the generation of the scallop genome, a single male from a family derived from the self-cross of a hermaphrodite was used for whole-genome sequencing and assembly (Wang et al., 2017).

The characterization of different bivalve genomes has revealed that these animals have more complex genomes. Among the most distinctive features of the bivalve genomes is their high level of polymorphism. The levels of sequence polymorphism in bivalves are higher than those reported for most studied animal genomes, such as humans, nematodes or shrimps (Hillier et al., 2008, Venter et al., 2001, Yu et al., 2014, Reich et al., 2003). For instance, the genome heterozygosity of a wild Pacific oyster individual was of 1.3% and 0.73% for an inbreed individual (Zhang et al., 2012). Whereas for the sequenced mussel species, the genome-wide heterozygosity rate was 1.24% for the deep-sea mussel and 2.02% for the Philippine horse mussel (Sun et al., 2017). Regarding repetitive elements, they were found to be similar in the Pacific oyster and scallop – representing 36% and 39% of the genome, respectively – and are dominated by tandem repeats. These values are not particularly high compared to other species such as the Atlantic salmon (58-60%; Lien et al. (2016)) or humans (66-69%; de Koning et al. (2011)). However, repetitive elements in the mussel genomes were comparatively higher, representing 48% of the genome of the deep sea-vent mussel and 62% of the Philippine horse mussel genome. Notably, in the Pacific oyster genome a significant fraction of repetitive elements is represented by mobile elements, which were found to be actively shaping genomic variation (Zhang et al., 2012).

1.1.6 Limitations of applying molecular markers in bivalves

As new genomic resources continue to expand for bivalve species, old challenges resurge. The limitations that researchers experienced in the early days of allozyme analysis are similar to those experienced in recent years with microsatellite (Rico et al., 2017) or SNP markers – i.e., high levels of SD. A concerning aspect of this issue is that the widely accepted explanation for SD is the linkage of markers with deleterious mutations, which are expressed near metamorphosis. For this high mutational load to exist in natural populations, with rates remaining relatively unaffected by natural selection, then they must be replenished continually. A high mutation rate would have profound consequences for the biological interpretation of genetic diversity and inheritance patterns, although it has rarely been considered in bivalve genetic studies. The reason for this may be that, despite being acknowledged as a relevant factor (Yu and Guo, 2003), they are still perceived to occur at negligible rates. Given that no study has directly assessed their frequency, the presumption of a particularly high mutation rate in bivalves remains to be tested. Thus, there is a knowledge gap in bivalve genetics – we are developing thousands of molecular markers but still do not know with confidence if they completely adhere to Mendelian rules. Altogether the frequent presence of SD and the uncertainty of high mutation rates, suggest that a deeper characterization of inheritance patterns would provide the connection required to understand why genetic markers in bivalves display such unique features. In the short-term, understanding the nature of the particularly complex genetic patterns of bivalves will become a necessity if markers are to be applied in instances in which a high precision is required (e.g., parental testing, genome-wide association studies, genomic selection).

Interest in bivalve genomics has emerged during the recent years, owing to the importance of these organisms as aquaculture resources and to their role in marine environmental science (Saavedra and Bachère, 2006). Enabled by NGS technologies, high-density molecular markers are increasingly becoming available for bivalve species, mainly driven by the aquaculture industry. Bivalve shellfish represent nearly 21 % (12.9 million tonnes) of world production and 10 % (US$12.8 billion) of world economic value of the aquaculture industry (FAO, 2010). In general, the bivalve farming industry relies heavily on the environment for spat (juveniles) supply. However, this spat varies year to year likely due to changes in environmental conditions (Robert and Gerard, 1999). Because of the uncertainty

of the long-term supply of seeds, and to ensure a sustainable growth of the industry, spat needs to start being produced in hatcheries. This shift in the production-cycle will bring significant opportunities for the genetic improvement of bivalve species via selective breeding programs. However, this could only be possible if accurate genetic information is being conveyed by genetic markers. To meet these challenges, basic research on primary aspects of genetic transmission has to be conducted.

1.2 Thesis aims and summary of approaches

In a previous study (Peñaloza, 2013) we evaluated the inheritance patters of SNP markers in two Chilean mussel families (*Mytilus chilensis*) by applying restriction-site associated DNA sequencing (RAD-Seq). The research revealed that heterozygote deficiency was, in fact, a strong, widespread genomic phenomenon; ~70% of the discovered markers showed SD. Further preliminary evidence of a high number of novel (*de novo*) SNP alleles was presented, as adult offspring carried genetic variants that were not observed in their parents. The substantial marker SD was rather consistent with previous findings. Nevertheless, the presence of novel alleles in adult individuals was unexpected, and provided the first empirical confirmation of predictions of high mutation rates in bivalves.

Extending upon the important research that has determined that SD in bivalves originate at the larval stage (Bierne et al., 1998, Launey and Hedgecock, 2001, Plough, 2016b, Plough and Hedgecock, 2011, Plough et al., 2016), and intrigued by the frequent presence of *de novo* alleles in F1 mussel progeny (Peñaloza, 2013), this dissertation revisits some long-standing, unusual genetic features of bivalves by making use of high-throughput sequencing technologies. By analyzing bivalve families at a genome-wide scale, an increased resolution of inheritance patterns is enabled, which allowed the investigation of the molecular basis of genetic transmission in greater detail than previous efforts that have focussed on limited numbers of genetic markers.

The overall aim of this dissertation is to provide insights into genomic patterns of SD and the potential origin of *de novo* alleles in bivalves. The general experimental approach involved creating families from pair-crosses of different bivalve species, genotyping individuals with high-throughput sequencing methods, and evaluating possible causes of the deviations of markers from Mendelian inheritance expectations. Two genome-wide

genotyping methods were used – a SNP chip (only used in Chapter 4) and the RAD-Seq method (used in Chapters 3, 4, 5 and 6). The first part of the research consisted of evaluating different *de novo* assembly and variant calling pipelines for the RAD-Seq data in bivalve species, which is of particular relevance when sequencing species with a complex genome (Chapter 3). Second, three Pacific oyster families were typed with two high-throughput genotyping technologies, a SNP array (~55k) and the RAD-Seq method. Levels of Mendelian errors were estimated for both technologies and compared, as a first approach towards assessing the presence of null alleles in widely used genotyping platforms. In addition, the high SNP marker density was used to characterize the genome-wide patterns of SD in the oyster genome (Chapter 4). Third, Mendelian errors detected in the RAD-Seq data of Chapter 4 were utilized to evaluate in detail the hypothesis that they were being caused by technical artifacts, namely null alleles. The approach incorporated marker locus read-depth to the analysis of marker inheritance patterns and evaluated the likelihood that patterns that fitted with the segregating of null alleles were responsible for Mendelian errors (Chapter 5). Finally, I aimed to determine the stage at which *de novo* mutations arise by sequencing gametes and larvae. Five Greenshell™ mussel families were created, and samples from adults, their respective gametes and offspring were sequenced. The sequencing of gametes provides an experimental way of examining the putative germline origin of *de novo* mutations. On the other hand, the sequencing of larval stages provides information on putative post-zygotic mutational events. A view of the main contributing factors to the genetic makeup of larval offspring provides insight into how the high levels of genetic variation are potentially maintained by bivalve populations (Chapter 6).

1.3 References

ASTORGA, M. P. 2014. Genetic considerations for mollusk production in aquaculture: current state of knowledge. *Frontiers in Genetics,* 5**,** 435.

BAI, Z., HAN, X., LUO, M., LIN, J., WANG, G. & LI, J. 2015. Constructing a microsatellite-based linkage map and identifying QTL for pearl quality traits in triangle pearl mussel (Hyriopsis cumingii). *Aquaculture,* 437**,** 102-110.

BEAUMONT, A. R. 1991. Genetic studies of laboratory reared mussels, Mytilus edulis: heterozygote deficiencies, heterozygosity and growth. *Biological Journal of the Linnean Society,* 44**,** 273-285.

BEAUMONT, A. R. & BEVERIDGE, C. M. 1984. *Electrophoretic survey of genetic variation in Pecten maximus, Chlamys opercularis, C. varia and C. distorta from the Irish Sea*.

BERG, P. 2006. Origins of the Human Genome Project: Why Sequence the Human Genome When 96% of It Is Junk? *American Journal of Human Genetics,* 79**,** 603-605.

BETTENCOURT, R., PINHEIRO, M., EGAS, C., GOMES, P., AFONSO, M., SHANK, T. & SANTOS, R. S. 2010. High-throughput sequencing and analysis of the gill tissue transcriptome from the deep-sea hydrothermal vent mussel Bathymodiolus azoricus. *BMC Genomics,* 11, 559.

BHARADWAJ, R. & YU, H. 2004. The spindle checkpoint, aneuploidy, and cancer. *Oncogene,* 23**,** 2016-2027.

BIERNE, N., LAUNEY, S., NACIRI-GRAVEN, Y. & BONHOMME, F. 1998. Early Effect of Inbreeding as Revealed by Microsatellite Analyses on Ostrea edulis Larvae. *Genetics,* 148**,** 1893-1906.

BIRKY, C. W., DEMKO, C. A., PERLMAN, P. S. & STRAUSBERG, R. 1978. Uniparental Inheritance of Mitochondrial Genes in Yeast: Dependence on Input Bias of Mitochondrial DNA and Preliminary Investigations of the Mechanism. *Genetics,* 89**,** 615-651.

BORSA, P., ZAINURI, M. & DELAY, B. 1991. Heterozygote deficiency and population structure in the bivalve Ruditapes decussatus. *Heredity,* 66**,** 1-8.

BOUILLY, K., LEITAO, A., MCCOMBIE, H. & LAPEGUE, S. 2003. Impact of atrazine on aneuploidy in pacific oysters, Crassostrea gigas. *Environmental Toxicology and Chemistry,* 22**,** 219-23.

BRETON, CAPT, GUERRA & STEWART. 2017. Sex determining mechanisms in bivalves. *Preprints.*

BRETON, S., BURGER, G., STEWART, D. T. & BLIER, P. U. 2006. Comparative analysis of gender-associated complete mitochondrial genomes in marine mussels (Mytilus spp.). *Genetics,* 172**,** 1107-1119.

BROWN, W. M., GEORGE, M. & WILSON, A. C. 1979. Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America,* 76**,** 1967-1971.

BURZYNSKI, A., ZBAWICKA, M., SKIBINSKI, D. O. & WENNE, R. 2006. Doubly uniparental inheritance is associated with high polymorphism for rearranged and recombinant control region haplotypes in Baltic Mytilus trossulus. *Genetics,* 174**,** 1081-1094.

CAO, L., KENCHINGTON, E. & ZOUROS, E. 2004. Differential segregation patterns of sperm mitochondria in embryos of the blue mussel (Mytilus edulis). *Genetics,* 166**,** 883-894.

CHAN, D. C. & SCHON, E. A. 2012. Eliminating mitochondrial DNA from sperm. *Developmental cell,* 22**,** 469-470.

CLARK, M. S., THORNE, M. A., VIEIRA, F. A., CARDOSO, J. C., POWER, D. M. & PECK, L. S. 2010. Insights into shell deposition in the Antarctic bivalve Laternula elliptica: gene

discovery in the mantle transcriptome using 454 pyrosequencing. *BMC Genomics,* 11**,** 362.
COPPE, A., BORTOLUZZI, S., MURARI, G., MARINO, I. A. M., ZANE, L. & PAPETTI, C. 2012. Sequencing and characterization of striped venus transcriptome expand resources for clam fishery genetics. *PLoS One,* 7, e44185.
CUNNINGHAM, C., HIKIMA, J.-I., JENNY, M., CHAPMAN, R., FANG, G.-C., SASKI, C., LUNDQVIST, M., WING, R., CUPIT, P., GROSS, P., WARR, G. & TOMKINS, J. 2006. New resources for marine genomics: bacterial artificial chromosome libraries for the Eastern and Pacific oysters (Crassostrea virginica and C. gigas). *Marine biotechnology (New York, N.Y.),* 8**,** 521-533.
DAVID, P., PERDIEU, M. A., PERNOT, A. F. & JARNE, P. 1997. FINE-GRAINED SPATIAL AND TEMPORAL POPULATION GENETIC STRUCTURE IN THE MARINE BIVALVE SPISULA OVALIS. *Evolution,* 51**,** 1318-1322.
DE KONING, A. P. J., GU, W., CASTOE, T. A., BATZER, M. A. & POLLOCK, D. D. 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genetics,* 7**,** e1002384.
DE LORGERIL, J., ZENAGUI, R., ROSA, R. D., PIQUEMAL, D. & BACHÈRE, E. 2011. Whole Transcriptome Profiling of Successful Immune Response to Vibrio Infections in the Oyster Crassostrea gigas by Digital Gene Expression Analysis. *PLOS ONE,* 6**,** e23142.
DEY, P. 2004. Aneuploidy and malignancy: an unsolved equation. *Journal of Clinical Pathology,* 57**,** 1245-1249.
DIEHL, W. J. & KOEHN, R. K. 1985. Multiple-locus heterozygosity, mortality, and growth in a cohort of Mytilus edulis. *Marine Biology,* 88**,** 265-271.
DIXON, D. R. 1982. Aneuploidy in mussel embryos (*Mytilus edulis* L.) originating from a polluted dock. *Marine Biology Letters,* 3**,** 155-161.
DOKIANAKIS, E. & LADOUKAKIS, E. D. 2014. Different degree of paternal mtDNA leakage between male and female progeny in interspecific Drosophila crosses. *Ecology and Evolution,* 4**,** 2633-2641.
DONDERO, F., BANNI, M., NEGRI, A., BOATTI, L., DAGNINO, A. & VIARENGO, A. 2011. Interactions of a pesticide/heavy metal mixture in marine bivalves: a transcriptomic assessment. *BMC Genomics,* 12**,** 195.
DUMBAULD, B. R., RUESINK, J. L. & RUMRILL, S. S. 2009. The ecological role of bivalve shellfish aquaculture in the estuarine environment: A review with application to oyster and clam culture in West Coast (USA) estuaries. *Aquaculture,* 290**,** 196-223.
DUPERRON, S., BERGIN, C., ZIELINSKI, F., BLAZEJAK, A., PERNTHALER, A., MCKINESS, Z., DECHAINE, E., CAVANAUGH, C. & DUBILIER, N. 2006. A dual symbiosis shared by two mussel species, Bathymodiolus azoricus and Bathymodiolus puteoserpentis (Bivalvia: Mytilidae), from hydrothermal vents along the northern Mid-Atlantic Ridge. *Environmental microbiology,* 8**,** 1441-1447.
FAO 2010. *The State of World Fisheries and Aquaculture,* Rome.
FAO 2014. Fishery and Aquaculture Statistics. *FAO yearbook.* Food and Agriculture Organization of the United Nations.
FISHER, C. & SKIBINSKI, D. O. F. 1990. Sex-Biased Mitochondrial DNA Heteroplasmy in the Marine Mussel Mytilus. *The Royal Society Proceedings: Biological Sciences,* 242**,** 149-156.
FOLTZ, D. W. 1986. NULL ALLELES AS A POSSIBLE CAUSE OF HETEROZYGOTE DEFICIENCIES IN THE OYSTER CRASSOSTREA VIRGINICA AND OTHER BIVALVES. *Evolution,* 40**,** 869-870.
GARRIDO-RAMOS, M. A., STEWART, D. T., SUTHERLAND, B. W. & ZOUROS, E. 1998. The distribution of male-transmitted and female-transmitted mitochondrial DNA types

in somatic tissues of blue mussels: Implications for the operation of doubly uniparental inheritance of mitochondrial DNA. *Genome,* 41**,** 818-824.
GENTRY, R. R., FROEHLICH, H. E., GRIMM, D., KAREIVA, P., PARKE, M., RUST, M., GAINES, S. D. & HALPERN, B. S. 2017. Mapping the global potential for marine aquaculture. *Nature Ecology & Evolution,* 1**,** 1317-1324.
GOSLING, E. 1992. *The mussel Mytilus: ecology, physiology, genetics and culture,* University of California, Elsevier.
GOSLING, E. M. & WILKINS, N. P. 1985. Genetics of settling cohorts of Mytilus edulis (L.): Preliminary observations. *Aquaculture,* 44**,** 115-123.
GUO, X., HE, Y., ZHANG, L., LELONG, C. & JOUAUX, A. 2015. Immune and stress responses in oysters with insights on adaptation. *Fish & Shellfish Immunology,* 46**,** 107-119.
GUTIERREZ, A. P., TURNER, F., GHARBI, K., TALBOT, R., LOWE, N. R., PEÑALOZA, C., MCCULLOUGH, M., PRODÖHL, P. A., BEAN, T. P. & HOUSTON, R. D. 2017. Development of a Medium Density Combined-Species SNP Array for Pacific and European Oysters (Crassostrea gigas and Ostrea edulis). *G3: Genes|Genomes|Genetics*, 7, 2209-2218.
HARRANG, E., LAPEGUE, S., MORGA, B. & BIERNE, N. 2013. A High Load of Non-neutral Amino-Acid Polymorphisms Explains High Protein Diversity Despite Moderate Effective Population Size in a Marine Bivalve With Sweepstakes Reproduction. *G3: Genes|Genomes|Genetics,* 3**,** 333-341.
HASSOLD, T., HALL, H. & HUNT, P. 2007. The origin of human aneuploidy: where we have been, where we are going. *Human Molecular Genetics,* 16**,** R203-R208.
HEDGECOCK, D., LIN, J., DECOLA, S., HAUDENSCHILD, C. D., MEYER, E. & MANAHAN, D. T. 2007. Transcriptomic analysis of growth heterosis in larval Pacific oysters (Crassostrea gigas). *Proceedings of the National Academy of Sciences of the United States of America*, 104, 2313-2318.
HEDGECOCK, D. & PUDOVKIN, A. I. 2011. Sweepstakes Reproductive Success in Highly Fecund Marine Fish and Shellfish: A Review and Commentary. *Bulletin of Marine Science,* 87**,** 971-1002.
HEDGECOCK, D., SHIN, G., GRACEY, A. Y., DEN BERG, D. V. & SAMANTA, M. P. 2015. Second-Generation Linkage Maps for the Pacific Oyster Crassostrea gigas Reveal Errors in Assembly of Genome Scaffolds. *G3: Genes|Genomes|Genetics,* 5**,** 2007-2019.
HELLER, J. 1993. Hermaphroditism in molluscs. *Biological Journal of the Linnean Society,* 48**,** 19-42.
HILLIER, L. W., MARTH, G. T., QUINLAN, A. R., DOOLING, D., FEWELL, G., BARNETT, D., FOX, P., GLASSCOCK, J. I., HICKENBOTHAM, M., HUANG, W., MAGRINI, V. J., RICHT, R. J., SANDER, S. N., STEWART, D. A., STROMBERG, M., TSUNG, E. F., WYLIE, T., SCHEDL, T., WILSON, R. K. & MARDIS, E. R. 2008. Whole-genome sequencing and variant discovery in C. elegans. *Nature Methods,* 5**,** 183-188.
HOEH, W. R., STEWART, D. T., SAAVEDRA, C., SUTHERLAND, B. W. & ZOUROS, E. 1997. Phylogenetic evidence for role-reversals of gender-associated mitochondrial DNA in Mytilus (Bivalvia: Mytilidae). *Molecular Biology and Evolution,* 14**,** 959-67.
HUBERT, S. & HEDGECOCK, D. 2004. Linkage maps of microsatellite DNA markers for the Pacific oyster Crassostrea gigas. *Genetics,* 168**,** 351-362.
JEFFARES, D. C., TOMICZEK, B., SOJO, V. & DOS REIS, M. 2015. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *Methods in Molecular Biology,* 1201**,** 65-90.
JIAO, W., FU, X., DOU, J., LI, H., SU, H., MAO, J., YU, Q., ZHANG, L., HU, X., HUANG, X., WANG, Y., WANG, S. & BAO, Z. 2014. High-Resolution Linkage and Quantitative

Trait Locus Mapping Aided by Genome Survey Sequencing: Building Up An Integrative Genomic Framework for a Bivalve Mollusc. *DNA Research,* 21**,** 85-101.
JONES, D., JERRY, D., FORÊT, S., KONOVALOV, D. & ZENGER, K. 2013. Genome-Wide SNP Validation and Mantle Tissue Transcriptome Analysis in the Silver-Lipped Pearl Oyster, Pinctada maxima. *Marine biotechnology (New York, N.Y.),* 15**,** 647-658.
KENCHINGTON, E., MACDONALD, B., CAO, L., TSAGKARAKIS, D. & ZOUROS, E. 2002. Genetics of Mother-Dependent Sex Ratio in Blue Mussels (Mytilus spp.) and Implications for Doubly Uniparental Inheritance of Mitochondrial DNA. *Genetics,* 161**,** 1579-1588.
KIM, S. K., OH, J. R., SHIM, W. J., LEE, D. H., YIM, U. H., HONG, S. H., SHIN, Y. B. & LEE, D. S. 2002. Geographical distribution and accumulation features of organochlorine residues in bivalves from coastal areas of South Korea. *Marine Pollution Bulletin,* 45**,** 268-279.
KOEHN, R. K. & GAFFNEY, P. M. 1984. Genetic heterozygosity and growth rate in Mytilus edulis. *Marine Biology,* 82**,** 1-7.
KORENBERG, J. R., CHEN, X. N., SCHIPPER, R., SUN, Z., GONSKY, R., GERWEHR, S., CARPENTER, N., DAUMER, C., DIGNAN, P. & DISTECHE, C. 1994. Down syndrome phenotypes: the consequences of chromosomal imbalance. *Proceedings of the National Academy of Sciences of the United States of America,* 91**,** 4997-5001.
KYRIAKOU, E., ZOUROS, E. & RODAKIS, G. C. 2010. The atypical presence of the paternal mitochondrial DNA in somatic tissues of male and female individuals of the blue mussel species Mytilus galloprovincialis. *BMC Research notes,* 3, 222.
LALLIAS, D., LAPÈGUE, S., HECQUET, C., BOUDRY, P. & BEAUMONT, A. R. 2007. AFLP-based genetic linkage maps of the blue mussel (Mytilus edulis). *Animal Genetics,* 38**,** 340–349.
LASSEN, H. H. & TURANO, F. J. 1978. Clinal variation and heterozygote deficit at the lap-locus in Mytilus edulis. *Marine Biology,* 49**,** 245-254.
LAUNEY, S. & HEDGECOCK, D. 2001. High genetic load in the Pacific oyster Crassostrea gigas. *Genetics,* 159**,** 255-265.
LEITÃO, A., BOUDRY, P., MCCOMBIE, H., GÉRARD, A. & THIRIOT-QUIÉVREUX, C. 2001. Experimental evidence for a genetic basis to differences in aneuploidy in the Pacific oyster (Crassostrea gigas). Aquatic Living Resources*,* 14**,** 233-237.
LI, L. & GUO, X. 2004. AFLP-Based Genetic Linkage Maps of the Pacific Oyster Crassostrea gigas Thunberg. *Marine Biotechnology,* 6**,** 26-36.
LIEN, S., KOOP, B. F., SANDVE, S. R., MILLER, J. R., KENT, M. P., NOME, T., HVIDSTEN, T. R., LEONG, J. S., MINKLEY, D. R., ZIMIN, A., GRAMMES, F., GROVE, H., GJUVSLAND, A., WALENZ, B., HERMANSEN, R. A., VON SCHALBURG, K., RONDEAU, E. B., DI GENOVA, A., SAMY, J. K. A., OLAV VIK, J., VIGELAND, M. D., CALER, L., GRIMHOLT, U., JENTOFT, S., INGE VÅGE, D., DE JONG, P., MOEN, T., BARANSKI, M., PALTI, Y., SMITH, D. R., YORKE, J. A., NEDERBRAGT, A. J., TOOMING-KLUNDERUD, A., JAKOBSEN, K. S., JIANG, X., FAN, D., HU, Y., LIBERLES, D. A., VIDAL, R., ITURRA, P., JONES, S. J. M., JONASSEN, I., MAASS, A., OMHOLT, S. W. & DAVIDSON, W. S. 2016. The Atlantic salmon genome provides insights into rediploidization. *Nature,* 533**,** 200-205.
LIU, Z., WANG, L., ZHOU, Z., SUN, Y., WANG, M., WANG, H., HOU, Z., GAO, D., GAO, Q. & SONG, L. 2016. The simple neuroendocrine-immune regulatory network in oyster Crassostrea gigas mediates complex functions. *Scientific Reports*, 6**,** 26396.
LONGWELL, A. C. & STILES, S. S. 1968. Fertilization and Completion of Meiosis in Spawned Eggs of the American Oyster, Crassostrea Virginica Gmelin. *Caryologia,* 21**,** 65-73.

MACLENNAN, M., CRICHTON, J. H., PLAYFOOT, C. J. & ADAMS, I. R. 2015. Oocyte development, meiosis and aneuploidy. *Seminars in cell & developmental biology,* 45**,** 68-76.

MALLET, A. L., ZOUROS, E., GARTNER-KEPKAY, K. E., FREEMAN, K. R. & DICKIE, L. M. 1985. Larval viability and heterozygote deficiency in populations of marine bivalves: evidence from pair matings of mussels. *Marine Biology,* 87**,** 165-172.

MARESCAUX, J. & VAN DONINCK, K. 2013. Using DNA barcoding to differentiate invasive Dreissena species (Mollusca, Bivalvia). *ZooKeys***,** 365, 235-244.

MARSHALL, DUSTIN J. & MORGAN, STEVEN G. 2011. Ecological and Evolutionary Consequences of Linked Life-History Stages in the Sea. *Current Biology,* 21**,** R718-R725.

MARTÍNEZ-LAGE, A., GONZÁLEZ-TIZÓN, A., AUSIÓ, J. & MÉNDEZ, J. 1997. Karyotypes and Ag-NORs of the mussels mytilus californianus and M. trossulus from the Pacific Canadian coast. *Aquaculture,* 153**,** 239-249.

MCGOLDRICK, D. J. & HEDGECOCK, D. 1997. Fixation, Segregation and Linkage of Allozyme Loci in Inbred Families of the Pacific Oyster Crassostrea gigas (Thunberg): Implications for the Causes of Inbreeding Depression. *Genetics,* 146**,** 321-334.

MEUWISSEN, T. H. E., HAYES, B. J. & GODDARD, M. E. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics,* 157**,** 1819-1829.

MURGARELLA, M., PUIU, D., NOVOA, B., FIGUERAS, A., POSADA, D. & CANCHAYA, C. 2016. A First Insight into the Genome of the Filter-Feeder Mussel Mytilus galloprovincialis. *PLOS ONE,* 11**,** e0151561.

NEWELL, R. 2004. Ecosystem influences of natural and cultivated populations of suspension-feeding bivalve molluscs: a review. *Journal of Shellfish Research,* 23**,** 51-61.

NIE, H., YAN, X., HUO, Z., JIANG, L., CHEN, P., LIU, H., DING, J. & YANG, F. 2017. Construction of a High-Density Genetic Map and Quantitative Trait Locus Mapping in the Manila clam Ruditapes philippinarum. *Scientific Reports,* 7**,** 229.

OBATA, M., SHIMIZU, M., SANO, N. & KOMARU, A. 2008. Maternal inheritance of mitochondrial DNA (mtDNA) in the Pacific oyster (Crassostrea gigas): a preliminary study using mtDNA sequence analysis with evidence of random distribution of MitoTracker-stained sperm mitochondria in fertilized eggs. *Zoological Science,* 25**,** 248-254.

PEÑALOZA, C. 2013. *Development and application of genomic tools to the genetic improvement of Atlantic salmon and Chilean mussels.* University of Edinburgh.

PLOUGH, L. V. 2012. Environmental stress increases selection against and dominance of deleterious mutations in inbred families of the Pacific oyster Crassostrea gigas. *Molecular Ecology,* 21**,** 3974-3987.

PLOUGH, L. V. 2016a. Genetic load in marine animals: a review. *Current Zoology,* 62**,** 567-579.

PLOUGH, L. V. 2016b. Fine-scale temporal analysis of genotype-dependent mortality at settlement in the Pacific oyster Crassostrea gigas. *bioRxiv*.

PLOUGH, L. V. & HEDGECOCK, D. 2011. Quantitative Trait Locus Analysis of Stage-Specific Inbreeding Depression in the Pacific Oyster Crassostrea gigas. *Genetics,* 189**,** 1473-1486.

PLOUGH, L. V., SHIN, G. & HEDGECOCK, D. 2016. Genetic inviability is a major driver of type III survivorship in experimental families of a highly fecund marine bivalve. *Molecular Ecology,* 25**,** 895-910.

QI, H., SONG, K., LI, C., WANG, W., LI, B., LI, L. & ZHANG, G. 2017. Construction and evaluation of a high-density SNP array for the Pacific oyster (Crassostrea gigas). *PLOS ONE,* 12**,** e0174007.

RAFF, R. A. 1987. Constraint, flexibility, and phylogenetic history in the evolution of direct development in sea urchins. *Developmental Biology,* 119**,** 6-19.

RAYMOND, M., XE, XE, NT, XF, L., R., THOMAS, F., XE, DERIC, ROUSSET, F., XE, OIS, DE, M., XFC, S, T., RENAUD, F., XE & OIS 1997. Heterozygote deficiency in the mussel Mytilus edulis species complex revisited. *Marine Ecology Progress Series,* 156**,** 225-237.

REICH, D. E., GABRIEL, S. B. & ALTSHULER, D. 2003. Quality and completeness of SNP databases. *Nature Genetics,* 33**,** 457-458.

RICO, C., CUESTA, J. A., DRAKE, P., MACPHERSON, E., BERNATCHEZ, L. & MARIE, A. D. 2017. Null alleles are ubiquitous at microsatellite loci in the Wedge Clam (Donax trunculus). *PeerJ,* 5**,** e3188.

RIVIERE, G., KLOPP, C., IBOUNIYAMINE, N., HUVET, A., BOUDRY, P. & FAVREL, P. 2015. GigaTON: an extensive publicly searchable database providing a new reference transcriptome in the pacific oyster Crassostrea gigas. *BMC Bioinformatics,* 16**,** 401.

ROBERT, R. & GERARD, A. 1999. Bivalve hatchery technology: The current situation for the Pacific oyster Crassostrea gigas and the scallop Pecten maximus in France. *Aquatic Living Resources*, 12, 121-130.

RODRIGUEZ-JUIZ, A. M., TORRADO, M. & MENDEZ, J. 1996. Genome-size variation in bivalve molluscs determined by flow cytometry. *Marine Biology,* 126**,** 489-497.

ROKAS A., LADOUKAKIS E. & ZOUROS E. 2003. Animal mitochondrial DNA recombination revisited. *Trends in Ecology and Evolution*. 18, 411-417.

RUPHIBASE. Available: http://compgen.bio.unipd.it/ruphibase/ [Accessed].

SAAVEDRA, C. & BACHÈRE, E. 2006. Bivalve genomics. *Aquaculture,* 256**,** 1-14.

SAAVEDRA, C., ZAPATA, C., GUERRA, A. & ALVAREZ, G. 1993. Allozyme variation in European populations of the oyster Ostrea edulis. *Marine Biology,* 115**,** 85-95.

SAKAMOTO, W., MIYAGISHIMA, S.-Y. & JARVIS, P. 2008. Chloroplast Biogenesis: Control of Plastid Development, Protein Import, Division and Inheritance. *The Arabidopsis Book***,** e0110.

SCHLUTER D., PRICE T.D. & L., R. 1991. Conflicting selection pressures and life history trade-offs. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 246**,** 11-17.

SEGARRA, A., MAUDUIT, F., FAURY, N., TRANCART, S., DÉGREMONT, L., TOURBIEZ, D., HAFFNER, P., BARBOSA-SOLOMIEU, V., PÉPIN, J.-F., TRAVERS, M.-A. & RENAULT, T. 2014. Dual transcriptomics of virus-host interactions: comparing two Pacific oyster families presenting contrasted susceptibility to ostreid herpesvirus 1. *BMC Genomics,* 15**,** 580.

SIMPSON, J. T. 2014. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics,* 30**,** 1228-1235.

SINGH, XU, J. & J., K. R. 2012. *Rapid Evolving Genes and Genetic Systems*.

SKIBINSKI, D. O., GALLAGHER, C. & BEYNON, C. M. 1994. Sex-limited mitochondrial DNA transmission in the marine mussel Mytilus edulis. *Genetics,* 138**,** 801-809.

SUN, J., ZHANG, Y., XU, T., ZHANG, Y., MU, H., ZHANG, Y., LAN, Y., FIELDS, C. J., HUI, J. H. L., ZHANG, W., LI, R., NONG, W., CHEUNG, F. K. M., QIU, J.-W. & QIAN, P.-Y. 2017. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nature Ecology & Evolution*, 1**,** 0121.

SUSSARELLU, R., FABIOUX, C., LE MOULLAC, G., FLEURY, E. & MORAGA, D. 2010. Transcriptomic response of the Pacific oyster Crassostrea gigas to hypoxia. *Marine Genomics,* 3**,** 133-143.

TAKEUCHI, T., KAWASHIMA, T., KOYANAGI, R., GYOJA, F., TANAKA, M. & IKUTA, T. 2012. Draft genome of the pearl oyster Pinctada fucata: a platform for understanding bivalve biology. *DNA Research,* 19, 117-130.

TAKEUCHI, T., KOYANAGI, R., GYOJA, F., KANDA, M., HISATA, K., FUJIE, M., GOTO, H., YAMASAKI, S., NAGAI, K., MORINO, Y., MIYAMOTO, H., ENDO, K., ENDO, H., NAGASAWA, H., KINOSHITA, S., ASAKAWA, S., WATABE, S., SATOH, N. & KAWASHIMA, T. 2016. Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. *Zoological Letters,* 2**,** 3.

TAYLOR, D. & INGVARSSON, P. 2003. Common Features of Segregation Distortion in Plants and Animals. *Genetica,* 117**,** 27-35.

TAYLOR, R. W. & TURNBULL, D. M. 2005. Mitochondrial DNA mutations in human disease. *Nature Reviews Genetics,* 6**,** 389-402.

THIRIOT-QUIÉVEUX, C. & AYRAUD, N. 1982. Les caryotypes de quelques espèces de bivalves et de gastéropodes marins. *Marine Biology,* 70**,** 165-172.

THIRIOT-QUIEVREUX, C. 2002. Review of the literature on bivalve cytogenetics in the last ten years. *Cahiers de Biologie Marine*, 43, 17-26.

THIRIOT-QUIÉVREUX, C., POGSON, G. H. & ZOUROS, E. 1992. Genetics of growth rate variation in bivalves: aneuploidy and heterozygosity effects in a Crassostrea gigas family. *Genome,* 35**,** 39-45.

TORO, J. E. & VERGARA, A. M. 1995. Evidence for Selection Against Heterozygotes: Post-Settlement Excess of Allozyme Homozygosity in a Cohort of the Chilean Oyster, Ostrea chilensis Philippi, 1845. *The Biological Bulletin,* 188**,** 117-119.

VENIER, P., DE PITTÀ, C., BERNANTE, F., VAROTTO, L., DE NARDI, B., BOVO, G., ROCH, P., NOVOA, B., FIGUERAS, A., PALLAVICINI, A. & LANFRANCHI, G. 2009. MytiBase: a knowledgebase of mussel (M. galloprovincialis) transcribed sequences. *BMC genomics,* 10**,** 72.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A., HOLT, R. A., GOCAYNE, J. D., AMANATIDES, P., BALLEW, R. M., HUSON, D. H., WORTMAN, J. R., ZHANG, Q., KODIRA, C. D., ZHENG, X. H., CHEN, L., SKUPSKI, M., SUBRAMANIAN, G., THOMAS, P. D., ZHANG, J., GABOR MIKLOS, G. L., NELSON, C., BRODER, S., CLARK, A. G., NADEAU, J., MCKUSICK, V. A., ZINDER, N., LEVINE, A. J., ROBERTS, R. J., SIMON, M., SLAYMAN, C., HUNKAPILLER, M., BOLANOS, R., DELCHER, A., DEW, I., FASULO, D., FLANIGAN, M., FLOREA, L., HALPERN, A., HANNENHALLI, S., KRAVITZ, S., LEVY, S., MOBARRY, C., REINERT, K., REMINGTON, K., ABU-THREIDEH, J., BEASLEY, E., BIDDICK, K., BONAZZI, V., BRANDON, R., CARGILL, M., CHANDRAMOULISWARAN, I., CHARLAB, R., CHATURVEDI, K., DENG, Z., FRANCESCO, V. D., DUNN, P., EILBECK, K., EVANGELISTA, C., GABRIELIAN, A. E., GAN, W., GE, W., GONG, F., GU, Z., GUAN, P., HEIMAN, T. J., HIGGINS, M. E., JI, R.-R., KE, Z., KETCHUM, K. A., LAI, Z., LEI, Y., LI, Z., LI, J., LIANG, Y., LIN, X., LU, F., MERKULOV, G. V., MILSHINA, N., MOORE, H. M., NAIK, A. K., NARAYAN, V. A., NEELAM, B., NUSSKERN, D., RUSCH, D. B., SALZBERG, S., SHAO, W., SHUE, B., SUN, J., WANG, Z. Y., WANG, A., WANG, X., WANG, J., WEI, M.-H., WIDES, R., XIAO, C., YAN, C., et al. 2001. The Sequence of the Human Genome. *Science,* 291**,** 1304-1351.

WANG, S., ZHANG, J., JIAO, W., LI, J., XUN, X., SUN, Y., GUO, X., HUAN, P., DONG, B., ZHANG, L., HU, X., SUN, X., WANG, J., ZHAO, C., WANG, Y., WANG, D., HUANG, X., WANG, R., LV, J., LI, Y., ZHANG, Z., LIU, B., LU, W., HUI, Y., LIANG, J., ZHOU, Z., HOU, R., LI, X.,

LIU, Y., LI, H., NING, X., LIN, Y., ZHAO, L., XING, Q., DOU, J., LI, Y., MAO, J., GUO, H., DOU, H., LI, T., MU, C., JIANG, W., FU, Q., FU, X., MIAO, Y., LIU, J., YU, Q., LI, R., LIAO, H., LI, X., KONG, Y., JIANG, Z., CHOURROUT, D., LI, R. & BAO, Z. 2017. Scallop genome provides insights into evolution of bilaterian karyotype and development. 1**,** 0120.

WHITLOCK, M. C. & DAVIS, B. 2001. Genetic Load. *eLS.* John Wiley & Sons, Ltd.

YAOHUA, S., HONG, K., XIMING, G., ZHIFENG, G., YAN, W. & AIMIN, W. 2009. Genetic linkage map of the pearl oyster, Pinctada martensii (Dunker). *Aquaculture Research,* 41**,** 35-44.

YARRA, T., GHARBI, K., BLAXTER, M., PECK, L. S. & CLARK, M. S. 2016. Characterization of the mantle transcriptome in bivalves: Pecten maximus, Mytilus edulis and Crassostrea gigas. *Marine Genomics,* 27**,** 9-15.

YU, Z. & GUO, X. 2003. Genetic linkage map of the eastern oyster Crassostrea virginica Gmelin. *The Biological bulletin,* 204**,** 327-338.

YU, Y., WEI, J., ZHANG, X., LIU, J., LIU, C., LI, F. & XIANG, J. 2014. SNP Discovery in the Transcriptome of White Pacific Shrimp Litopenaeus vannamei by Next Generation Sequencing. *PLOS ONE,* 9**,** e87218.

YUAN, T., HE, M., HUANG, L. & HU, J. 2010. Genetic Linkage Maps of the Noble Scallop Chlamys nobilis Reeve Based on Aflp and Microsatellite Markers. *Journal of Shellfish Research,* 29**,** 55-62.

ZHANG, G., FANG, X., GUO, X., LI, L., LUO, R. & XU, F. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature,* 490.

ZHANG, L., BAO, Z., CHENG, J., LI, H., HUANG, X., WANG, S., ZHANG, C. & HU, J. 2007. Fosmid Library Construction and Initial Analysis of End Sequences in Zhikong Scallop (Chlamys farreri). *Marine Biotechnology,* 9**,** 606-612.

ZHANG, N., XU, F. & GUO, X. 2014. Genomic Analysis of the Pacific Oyster (Crassostrea gigas) Reveals Possible Conservation of Vertebrate Sex Determination in a Mollusc. *G3: Genes|Genomes|Genetics,* 4**,** 2207-2217.

ZHAO, X., YU, H., KONG, L. & LI, Q. 2012. Transcriptomic responses to salinity stress in the Pacific oyster Crassostrea gigas. *PLoS One,* 7, e46244.

ZOUROS, E., FREEMAN, K. R., BALL, A. O. & POGSON, G. H. 1992. Direct evidence for extensive paternal mitochondrial DNA inheritance in the marine mussel Mytilus. *Nature,* 359**,** 412-414.

ZOUROS, E., OBERHAUSER BALL, A., SAAVEDRA, C. & FREEMAN, K. R. 1994. An unusual type of mitochondrial DNA inheritance in the blue mussel Mytilus. *Proceedings of the National Academy of Sciences of the United States of America,* 91**,** 7463-7467.

ZOUROS, E., SINGH, S. M., FOLTZ, D. W. & MALLET, A. L. 1983. Post-settlement viability in the American oyster (Crassostrea virginica): an overdominant phenotype. *Genetics Research,* 41**,** 259-270.

ZOUROS, E., SINGH, S. M. & MILES, H. E. 1980. Growth Rate in Oysters: An Overdominant Phenotype and Its Possible Explanations. *Evolution,* 34**,** 856-867.

ZOUROS, E., THIRIOT-QUIEVREUX, C. & KOTOULAS, G. 2009. The negative correlation between somatic aneuploidy and growth in the oyster Crassostrea gigas and implications for the effects of induced polyploidization. *Genetical Research,* 68**,** 109-116.

CHAPTER 2

General Material and Methods

2.1 Material and Methods

2.1.1 Larval rearing and sampling

In the following Chapters of this thesis we study the genetics of bivalves by focusing on three species: the Greenshell™ mussel *Perna canaliculus*, the Blue mussel *Mytilus edulis* and the Pacific oyster *Crassostrea gigas*. The process of family creation and rearing for each species is described in the following section.

2.1.1.1 Greenshell™ (*Perna canaliculus*) mussel families

Two full-sibling and three half-sibling Greenshell™ mussel families were created from single bi-parental crosses at the Cawthron Institute, New Zealand, in October 2016. The origin of the sires used as broodstock was different from that of the dams. The sires derived from a sample of 50 wild-caught Greenshell™ mussels that were cleaned from epibionts and transferred to the laboratory for conditioning; whereas the dams belonged to a commercial growth-selected line. Spawning of mussels was induced via thermal stress in a communal tank. Males and females were identified at the time of spawning by the color of their gametes and were separated from the rest (Figure 1). Gametes from each individual were collected successively in plastic containers. To ensure the purity of the gametes, only the third collection of gametes was used for fertilization (the first and second collections were discarded).

Figure 1. **Individually spawned Greenshell™ mussels**.

To account for tank effects, each family was created in replicate. That is, eggs and sperm from the same parents were combined in two different 1-L beakers containing treated (filtered and UV irradiated) sea water that had been aged for a day. Spawning was induced by thermal shock. A total of 2-3 million eggs per family replicate were fertilized at a 1:200 egg to sperm ratio. Sperm and egg concentration was estimated from microscope counts of subsamples in a haemocytometer and a Sedgewick-rafter counting chamber, respectively. After 30 minutes of intermittent mixing with a plastic plunger, the developing embryos were transferred to 170-L static incubation tanks (Figure 2). All tanks contained 1 micron filtered sea water that was maintained at 18^{o}C with constant aeration. To minimize the chance of cross-contamination of individuals between families, we incremented the physical distance between incubation tanks by interspacing empty tanks. At three days post-fertilization, the larval offspring of the incubation tanks were drained through a screen, rinsed into 1-L beakers, and the number of larvae estimated. From each incubation tank (x2 per family due to replication), approximately 750,000 larvae were used to stock, again in replicate, the tanks of the Cawthron Ultra-Density Larval (CUDL) rearing system (Ragg et al., 2010) (Figure 2). The CUDL rearing system is comprised of an array of 2.5-L acrylic tanks designed to raise veliger larva at a high density (in this study ~ 300 larvae/ml). Each tank is connected to a feeding-line that provides a pre-mixed algal suspension (see below) at a constant rate. In addition, filtered air is injected into each tank with a glass

dropper, as a means of inducing gentle agitation. In this new rearing system, again we arranged the tanks to maximize the distance between different replicates and families.

The feeding scheme for mussel larvae consisted of a mixed diet of *Chaetoceros calcitrans* (*C. calc*) and *Isochrysis galbana* (*I. galb*). From day 3 until day 10 of cultivation the offspring were fed *ad libitum* a 2:1 *C. calc* to *I. galb* algae mixture at 10,000 cells/ml. As larvae grew larger, we increased the number of algal cells to 50,000 cells/ml. At day 12 post-fertilization, and because larval survival in our experiment was higher than expected, we had to reduce larval density in order to ensure the long-time survival of the offspring. Larger individuals were kept to continue with the culture (i.e., those retained on a 175-micron mesh), whereas smaller individuals were discarded. To account for the potential effect of this artificial size-selection on the estimates of offspring genotype proportions, both size categories were sampled for genetic analysis. Once all family replicates had individuals that showed a settling behavior (e.g., distinct eyespots and exploratory behavior; Lutz and Kennish (1992)), we divided the larval culture by size into two categories: (i) early settlers, represented by the largest larvae, and (ii) late settlers, represented by the smallest size class. Briefly, larval offspring from all families were screened through a 175-micron mesh: the larvae that were retained on the screen were relocated to new 2.5-L CUDLs (early settlers), whereas those that passed were returned to the same CUDL (late settlers). Larvae were provided with a rope as settling substrate and were fed *ad libitum* an algae mixture of *I.galb* and *C.cal* in a 2:1 ratio. The settled offspring were allowed to grow for 2 additional months, then samples from all families and replicates were taken for future genetic analysis.

Figure 2. Rearing systems used for growing the Greenshell™ mussel progeny. To the left, interspaced empty 170-L incubation tanks used in the first phase of larval rearing. To the right, CUDL rearing system (2.5-L) used for rearing larvae from 72 hours post-fertilization to the termination of the experiment at 2 months after settlement.

2.1.1.1.1 Sampling

Five parental tissues were sampled and preserved in absolute ethanol: gonad, adductor muscle, foot, gill and mantle. Because one of the aims of creating the Greenshell™ mussel families was to detect germline *de novo* mutations, we collected pools of sperm and eggs from the mussel parents. Pools of gametes were snap frozen in liquid nitrogen and preserved at -80^{o}C until DNA extraction. The Greenshell™ mussel has a 17-day larval stage at 18^{o}C. Mussels experience three conspicuous morphological transitions during post-embryonic development: from trochophore to veliger; from veliger to pediveligers; and from pediveligers to settled mussels (Gosling, 1992). For the temporal analysis aimed at detecting significant changes in allele frequency during larval development, samples of 5,000-10,000 larvae were taken from each family at ~20 hpf (trochophore), ~70 hpf (early veliger), 4 days post fertilization (dpf) (veliger), 8 dpf (veliger), and 12 dpf (late veliger). At

day 12 post-fertilization, samples were taken from both early and late settlers (defined above), corresponding to the group of large and small size larvae, respectively. Additional to the larval stage, offspring were also sampled at earlier stages of their benthic life, just after settlement. Samples were taken from settled individuals at day 15 post-settlement (=30 dpf) and at day 180 (=195 dpf). Samples of settled mussel offspring were collected in 50 ml Falcon tubes and preserved in absolute ethanol at a 1/3 ratio.

DNA was extracted from the pools of offspring using protocols developed in this thesis (section below). Unfortunately, day 12 samples from the late settler category (i.e., small size larvae) experienced unsupervised de-frosting from -80^{o}C to -20^{o}C. The impact of this change in temperature led to highly fragmented DNA extractions that were not suitable for the NGS technology used in this thesis (RAD-Seq). These samples were therefore discarded from the analysis. Settled offspring samples were not included in this thesis, but are currently stored at the Cawthron Institute, New Zealand for future studies.

2.1.1.2 Blue mussel (*Mytilus edulis*) families

In 2014, eighteen half-sibling and six full-sibling British blue mussel families (total n = 24 families) were created at the Centre for Environment, Fisheries and Aquaculture Science (Cefas), Weymouth, UK. We used a half-sib family design to evaluate whether the patterns of inheritance in mussel progeny are influenced by the parent of origin. To account for rearing-environment effects, three families were randomly chosen to be grown in replicate. To evaluate the potential risk of gamete cross-contamination during artificial fertilization, a single control family was created one week before establishing the rest of the families used in this experiment.

Adult mussels were collected from different wild and farmed populations across the UK (Fig. 3) and transferred to the laboratory located in Weymouth to be conditioned for spawning.

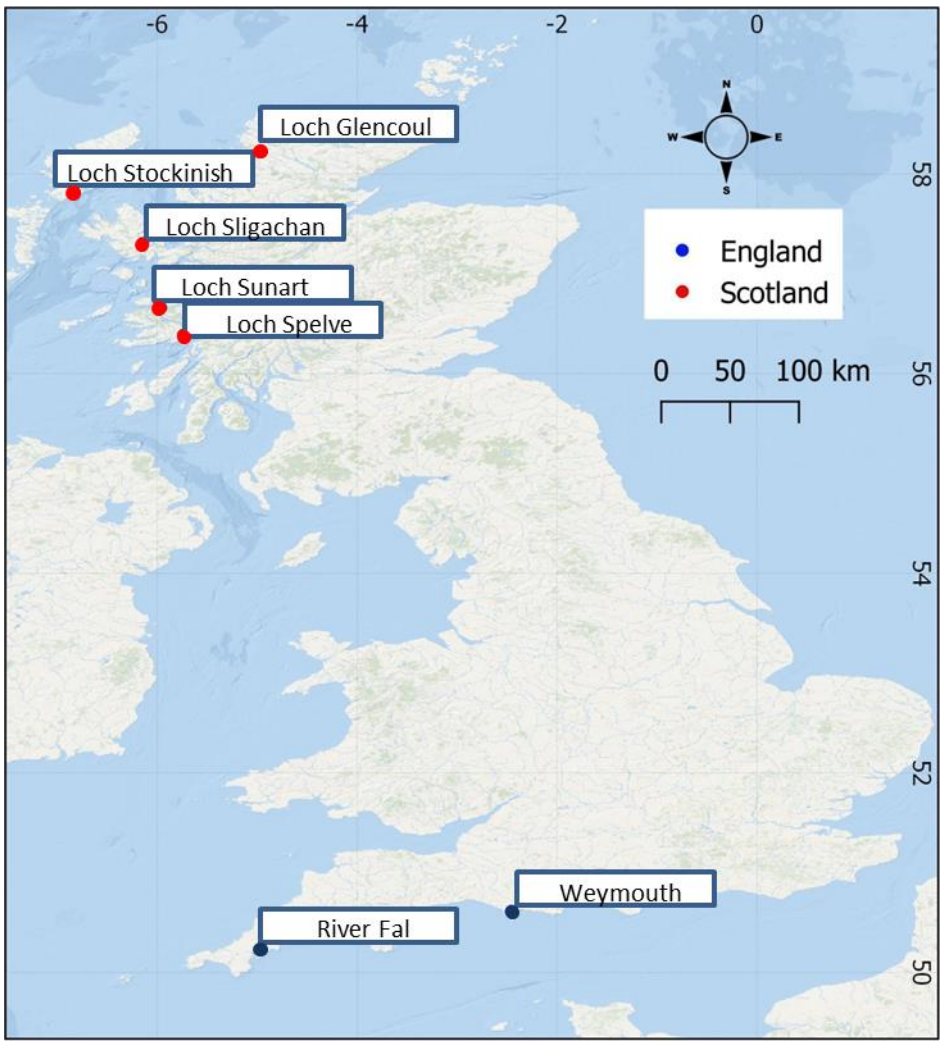


Figure 3. Map of the UK showing locations where the parent stock was sourced.

Ripe mussels were kept either in flow-through tanks with treated seawater (UV-sterilized and filtered at 0.2 microns) or in moist conditions until the experiment started. Experiments were carried out in an insulated cold room set at 17^{o}C. Spawning was induced by thermal shock: mussels were maintained for 1 hour on ice and then placed in individual beakers containing 400ml of heated sea water (24-26^{o}C). They were monitored at 20 min intervals, and once a spawning individual was detected the beaker was covered with parafilm, in order to prevent cross-contamination of gametes between beakers. The sex of spawning mussels was determined visually – females release a tightly formed cluster of bright orange eggs, whereas males release a milky white fluid. Sperm and egg concentration was estimated from microscope counts of subsamples in a hemocytometer and a Sedgewick-

rafter counting chamber, respectively. Crosses were performed by fertilizing 1000 eggs/ml at a sperm:egg ratio of 50:1 in 1-L beakers. The percentage of successful fertilization was assessed by the appearance of the first polar body. Blue mussel larval offspring were grown in a static system, at 17^{o}C in filtered and UV-irradiated seawater (Figure 4). Fertilized eggs were allowed to develop undisturbed until the veliger larval stage was reached, approximately 72hrs post-fertilization. Subsequently, sea water was changed every two days. At each water exchange, larvae were carefully collected in a 0.2-micron mesh to remove debris and exogenous species (e.g., rotifers). Dedicated, family-specific, materials were used throughout the experiment to prevent cross-contamination of progeny during the rearing procedure.

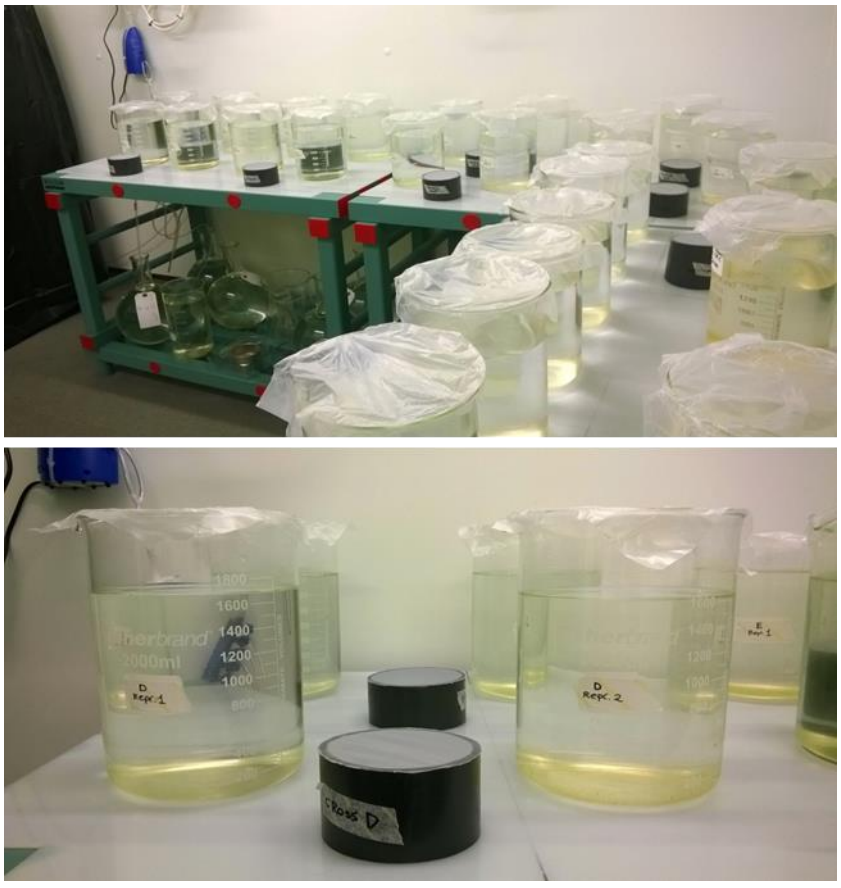


Figure 4. Blue mussel larval offspring were reared in a static culture system.

The feeding regime of mussel larvae followed that provided by Helm and Bourne (2004) for the Pacific oyster *Crassostrea gigas*. Different larval stages have different dietary requirements, reflecting their differing metabolic needs. Larvae were initially fed with a combination of the diatom *Chaetoceros muelleri* and the flagellate *Isochrysis galbana*. A third phytoplankton species, *Pavlova lutherii*, was incorporated into the diet when the larvae reached a length of 200µM. Mussel offspring were maintained at a culture density of 10 larvae mL^{-1}. Normal larval development (Figure 5) was monitored daily by observing 1 ml aliquots of the culture under a stereoscopic microscope. Once the larvae metamorphosed into pediveligers and showed signatures of competency, they were provided with a surface for settlement by placing a rope at the bottom of the beaker.

2.1.1.2.1 Sampling

Different tissue (gill, mantle, gonad, and foot) from the pool of sires and dams used as breeding stock were sampled and preserved in absolute ethanol. In addition, at the moment of spawning, a suspension of sperm and eggs from each parent were collected and cryopreserved in 5% DMSO. The F1 progeny from all mussel families were sampled at two larval stages: trochophore at 15 hpf, and the veliger stage at 72 hpf. Larval samples were gradually frozen to -80^{o}C in a mixture of 10% DMSO. The preservation method for gametes and larval offspring was chosen based on the preliminary evaluation of different cryopreservation protocols.

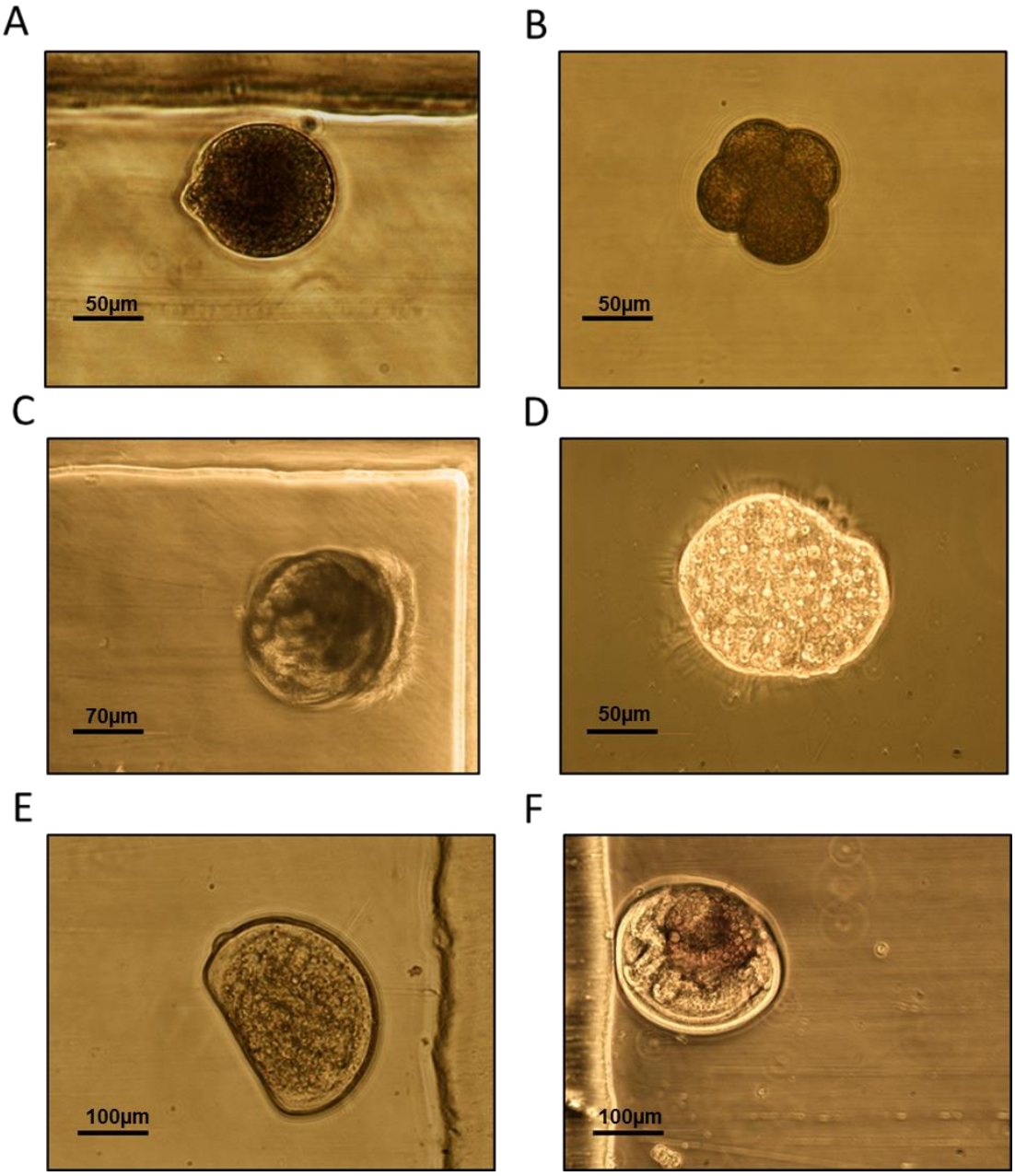


Figure 5. Embryonic and larval developmental stages of mussels. (A) Fertilised egg with the first polar body, (B) Embryo at the four-cell stage, (C) Transitional stage between the Trochophora and Veliger, (D) Trochophora (~16 hpf), (E) Veliger D-shaped larvae (~72 hpf) and (F) Pediveliger larval stage.

2.1.1.3 Pacific oyster *Crassostrea gigas* families

Three full-sibling families were created in 2016 by staff at Cefas, Weymouth, UK. Two families shared the same dam. To generate these families, adults containing ripe gametes were obtained from Guernsey Sea Farms (The UK). The method used for gamete release consisted of opening the oysters by breaking of the hinge, inserting a clean glass Pasteur pipette into the gonad to a depth of 1 – 2 mm, and aspirating the gonadal tissue into clean, 0.2-micron-filtered seawater irradiated with UV. The presence of sperm or eggs was determined by assessing sub-samples under a stereoscope microscope at 100x magnification. Sperm and eggs were filtered through 60 and 90 µm mesh sieves respectively to remove tissue debris. Eggs were diluted with seawater to a concentration of 3000 – 6000 eggs/ml. Approximately 2-3 ml of sperm was added per liter of eggs and stirred every 20 – 30 minutes. Larvae were reared in a static system, with constant aeration, at 25°C in 25 ppt seawater. Larval offspring were fed with a mixture of *Chaetoceros muelleri*, *Isochrysis galbana* and *Pavlova lutherii*, following guidelines proposed for larvae at different stages of development (Helm and Bourne, 2004). Once the oysters had grown to over 100µm in diameter, *C. muelleri* was gradually replaced with *Tetraselmis* sp., as the food requirements of larvae increase. After 4 weeks, ready-to-settle oysters were transferred to an upwelling system with flow-through water ($22^{o}C$ +- $2^{o}C$) and grown for 8 months until the termination of the experiment.

2.1.1.3.1 Sampling

Mantle tissue and gamete samples from each parent were collected at the time of fertilization. Tissue samples were preserved in absolute ethanol. Whole individual offspring were frozen and transferred to the Roslin Institute for genetic analysis.

2.1.2 DNA extraction methods

The isolation of high-quality genomic DNA is a key prerequisite for the high-throughput sequencing technique used in this thesis, RAD-Seq. The same DNA extraction procedure is unlikely to be optimal for DNA extraction from different biological sources, as expected from different sample composition. As follows, we describe the DNA extraction methods that were utilized for extracting genomic DNA of high quality from the three different sample types used in this study.

2.1.2.1 Adults and juveniles

Genomic DNA was extracted from all mussel samples using a CTAB (cetyl trimethylammonium bromide) method modified from Richards et al. (2013), as follows:

1. Slice approximately 200 mg of tissue with a clean scalpel. Place tissue in a 1.5 ml Eppendorf tube containing 400 µl of CTAB extraction solution (3% CTAB, 100 mM Tris-HCl, pH 7.5, 25 mM EDTA, pH 8, 2 m NaCl).
2. Add 5 µl proteinase K (20mg/ml) and mix by inverting tube 8-10 times.
3. Incubate at 65°C until lysed (~60 minutes or longer, depending upon tissue thickness), and mix by inverting occasionally.
4. Add 2ul of RNAse (10mg/ml) and leave at 37°C for 30 minutes.
5. Add 400 µl chloroform/isoamyl alcohol (24:1). Shake gently to emulsify. Leave at room temperature for 2 minutes and shake again.
6. Spin lysate at 13 000 rpm for 2 minutes in a bench top centrifuge, and transfer the upper layer to a clean 1.5 ml Eppendorf tube. Add 1 volume of chloroform/isoamyl alcohol (24:1).
7. Spin at 13 000 rpm for 2 minutes and transfer the upper layer to a new 1.5ml tube containing 900 µl CTAB dilution solution (1% CTAB, 50 mM Tris-HCl, pH 7.5, 10 mM EDTA, pH 8). Mix by inverting gently.
8. Spin at 13 000 rpm for 10 minutes.
9. Remove the supernatant with a micropipette, and then add 1000 µl of 0.4 M NaCl in TE, inverting tube to wash the pellet.
10. Spin again at 13 000 rpm for 5 minutes, and remove supernatant.

11. Add 300 µl 1.42 M NaCl in TE. This dissociates CTAB from DNA. Invert tube until pellet appears transparent.
12. Add 600 µl ethanol previously left at -20°C. Invert by hand and leave the DNA to precipitate overnight at -20°C.
13. Centrifuge at full speed for 8 minutes. Remove supernatant. Dry at room temperature, or in a heat block for 37 °C.
14. Suspend in 50 µl of buffer EB.

Genomic DNA from the Pacific oyster samples (adults + ~2-month-old offspring) was extracted with the REALPURE® Genomic DNA extraction kit (REAL RBMEG03, Durviz, Spain), following the manufacturer's instructions.

2.1.2.2 Larvae

Two types of larvae were sampled in this study: trochophore and veligers. Each larval stage posed unique challenges for DNA extraction methods. The main limitation for extracting DNA from the trochophore is that they are positively buoyant (personal observation). Therefore, they were not easily separated from the sea-water after collection. Additionally, they appeared to be very fragile and were prone to cell disintegration at high-speed centrifugation. Regarding the veliger stage, the main complication was that at this stage bivalve species develop a calcium carbonate shell, which required being broken before DNA was released by chemicals. Taking these issues into consideration, stage-specific extractions methods were developed to maximize DNA yield and quality.

DNA was extracted from the pooled trochophore larvae by successive centrifugation steps. Three vials of trochophore (~5,000 larvae each) were concentrated by centrifuging at low speed, then removing the supernatant and adding another vial on top, then centrifuged again, and so on, until the three vials were completed. Once the sample was concentrated, DNA was extracted from the pellet of larvae using the cell DNA extraction protocol from the DNAeasy Blood and Tissue kit (Qiagen).

To extract DNA from the pooled, shelled veliger larvae we crushed each sample three times on dry ice with a sterilized plastic pestle. The sample was left incubating with ATL buffer at 56^{o}C for 1 hour; additional mechanic disruption with plastic pestles (20 min intervals) is required for the larger size larvae (e.g., day 12 Greenshell™ mussel larvae). After incubation, the tubes were spun at 100g x 2 min to concentrate the shells in the bottom. The supernatant was transferred to a clean 1.5 Eppendorf tube. DNA was extracted from this supernatant following the manufacturer's instructions provided in the cell DNA extraction protocol from the DNAeasy Blood and Tissue kit (Qiagen).

2.1.2.3 Gametes

Total genomic DNA was extracted from bivalve mollusc gametes (sperm and eggs) following an online user developed protocol based on the DNAeasy Blood and Tissue kit (Purification of total DNA from animal sperm using the DNeasy® Blood & Tissue Kit; protocol 2). Although this protocol is standardized for DNA extraction from sperm, it was also utilized for the isolation of DNA from the oocytes, given that it performed comparatively better than other protocols tested (e.g., CTAB method).

2.1.3 RAD sequencing

Restriction Associated DNA sequencing (RAD-Seq) (Baird et al., 2008, Miller et al., 2007) was the main genotyping technology used in this thesis. The same library preparation method was utilized for the three species, with slight variation in the amount of input DNA.

The method for library preparation was as follows: DNA extractions were assessed by 1% agarose gel electrophoresis and 260/280 and 260/230 Nanodrop measurements (260/280>1.8, 260/230>2.0). Quantification of the DNA extraction with PicoGreen fluorometry (ds BR assay), samples were diluted to a working concentration of 25ng/ul. Each sample was digested with a SbfI high-fidelity restriction enzyme from the New England Biolabs (NEB) at 37°C for 60 min. The enzyme was heat inactivated for 20 min at 65°C. For ligation of P1 adaptors to the SbfI restriction site, 10nM of SbfI-P1 Adapters containing a 5 bp Molecular Identifier (MID) sequence were added to each sample. The ligation reaction was incubated overnight at 16°C. After heat inactivation at 65^{o}C for 20 min, samples were multiplexed so that each pool contained six samples. Pools of DNA were sheared to a size range from ~200 to 500bp using a Bioruptor® (Diagenode). Samples were cooled down on ice for 10 min before sonication. The settings for sonication were: 9s on followed by 30s off for 8 cycles. To promote a homogeneous shearing of the fragments, after the 4th cycle, sample pools were removed, spun down and gently mixed with a micropipette. The sheared, pooled samples were then purified using AMPure beads (Agencourt) at a 1.8:1 ratio of beads to DNA. Fragments ranging from ~300-500 bp were size-selected by agarose gel electrophoresis. A physical gap between libraries (i.e., pooled samples) was made in the gel to avoid cross-well contamination. After purification from the gel slice, libraries were end-repaired using the New England Biolabs (NEB) sample preparation kit (using an incubation time of 30 min). Following end-repair, libraries were purified using a 1.8:1 ratio of AMPure Beads to DNA. A single dA-tail was added to the 3' end of each blunted, phosphorylated DNA strand using Klenow Fragment DNA Polymerase (3'->5' exo-) from the NEB sample preparation kit (incubation time of 30 min at 37^{o}C). At the end of the incubation, the libraries were again purified using a 1.8:1 (Beads/DNA) ratio. The fragments were ligated to a P2 adapter (at 10 nM) with the T4 DNA Ligase. The ligation reaction was cleaned with a 0.9:1 ratio of beads to DNA, and eluted in 50 µL of buffer EB. Ten independent PCR reactions of 3ul each were performed to reduce PCR bias, using the Phusion HF Master Mix (NEB). The Phusion PCR settings followed product guidelines for a total of 18 cycles. PCR-enriched libraries were cleaned with 0.9:1 (beads/DNA) ratio, electrophoresed on a 1.1 % gel and excised between 300–700 bp. Fragments were purified from the gel slice using a MinElute Gel Extraction Kit (Qiagen) and eluted in 20 µL of buffer EB. All sub-libraries (i.e., ligated and PCR amplified pools of six samples) were mixed into a single master pool. The quality of the RAD-seq library was assessed using the 2200

TapeStation (Agilent Technologies). Libraries were diluted to a concentration of 10nM and were sequenced on a single lane of either a Illumina HiSeq 2000 or an Illumina HiSeq 2500 instrument by the Edinburgh Genomics Sequencing Facility, University of Edinburgh (https://genomics.ed.ac.uk/).

2.2 References

BAIRD, N. A., ETTER, P. D., ATWOOD, T. S., CURREY, M. C., SHIVER, A. L. & LEWIS, Z. A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One,* 3, e3376.

GOSLING, E. 1992. *The mussel Mytilus: ecology, physiology, genetics and culture,* University of California, Elsevier.

HELM, M. & BOURNE, N. 2004. Hatchery culture of bivalves. *FAO Fisheries technical paper* [Online], 471.

LUTZ, R. A. & KENNISH, M. J. 1992. Ecology and morphology of larval and early postlarval mussels. *The mussel Mytilus: ecology, physiology, genetics and culture. Elsevier, Amsterdam, The Netherlands***,** 53-85.

MILLER, M. R., DUNHAM, J. P., AMORES, A., CRESKO, W. A. & JOHNSON, E. A. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research,* 17**,** 240-248.

RAGG, N. L. C., KING, N., WATTS, E. & MORRISH, J. 2010. Optimising the delivery of the key dietary diatom Chaetoceros calcitrans to intensively cultured Greenshell™ mussel larvae, Perna canaliculus. *Aquaculture,* 306**,** 270-280.

RICHARDS, P. M., LIU, M. M., LOWE, N., DAVEY, J. W., BLAXTER, M. L. & DAVISON, A. 2013. RAD-Seq derived markers flank the shell colour and banding loci of the Cepaea nemoralis supergene. *Molecular Ecology,* 22**,** 3077-3089.

CHAPTER 3

Evaluation of *de novo* assembly and variant calling methods for bivalve RAD-Seq data

3.1 Introduction

Restriction site-associated DNA sequencing (RAD-Seq) has been widely used to simultaneously discover and genotype genome-wide polymorphisms in non-model organisms. In RAD-Seq the complexity of the genome under study is reduced through the enrichment of DNA sequences contiguous to restriction enzyme cut sites. This strategy enables the sub-sampling of a genome at putative homologous loci across many individuals for DNA sequencing, and large-scale SNP discovery and genotyping. Given that no prior genomic information of the organisms under study is required, and dozens of individuals may be sequenced simultaneously to potentially discover and genotype thousands to tens of thousands SNPs, thus lowering sequencing costs, RAD-Seq has become a staple technique for genetic studies in non-model organisms. Since the publication of the novel RAD-Seq approach (Baird et al., 2008), several variations have emerged (e.g. ddRAD Peterson et al. (2012); 2b-RAD Wang et al. (2012); and ezRAD Toonen et al. (2013), among others), each with their own benefits and drawbacks (Andrews et al., 2014), which has further expanded the versatility of this set of techniques. As powerful tools for genome analysis, the family of RAD-Seq methods have been used to obtain genetic information from a variety of species for a range of applications, including genetic mapping (Chutimanitsakun et al., 2011, Gonen et al., 2014), phylogenetic inference (Takahashi et al., 2014), and the identification of genomic regions responsible for adaptation and speciation (Tariel et al., 2016).

Most RAD-Seq studies aim to generate large SNP datasets across families, populations or species. Although RAD-Seq data has proven to be a reliable sequencing strategy (Catchen et al. (2017), although see Lowry et al. (2017)), as with any technique, errors may arise during the data generation process, which in turn may bias relevant biological estimates based on the inferred SNPs and genotypes (e.g., inbreeding coefficients or estimates of population structure). From the quality of the DNA source (Graham et al., 2015) to the size distribution of sheared DNA fragments (Davey et al., 2013), a range of potential sources of technical noise have been identified for the RAD-Seq laboratory protocol. However, an additional step that has received little attention, yet has been shown to significantly affect the amount and quality of the variants identified is RAD-Seq assembly, specifically *de novo* assembly (Pante et al., 2015, Mastretta-Yanes et al., 2015). For species that lack a reference genome, *de novo* assembly of short reads into RAD loci (i.e., DNA sequences flanking a restriction

site) is of crucial importance. The purpose of *de novo* assembly is to accurately reconstruct a complete set of loci within and across individuals for variant discovery. The process of loci reconstruction depends on the assembly parameter values, the choice of which will depend on: (i) the biology of the species under study, specifically their inherent level of polymorphism; (ii) the experimental design, mainly how many samples were simultaneously sequenced (multiplexed), given this will affect sample coverage; and (iii) externalities, such as poor quality DNA. Consequently, *de novo* parameter settings should be tailored to a RAD-Seq dataset based on a previous parameter optimization step (Paris et al., 2017, Mastretta-Yanes et al., 2015).

A few optimization approaches have been developed for *de novo* assembly of short reads obtained from the genomes of less studied species. Ilut et al. (2014) suggest assembling datasets at a series of incremental clustering threshold values and locating the value at which the number of single haplotype clusters (putative homozygous loci) and two haplotype clusters (putative heterozygous loci) asymptote. This value represents the limit beyond which the over-splitting of loci is negligible. Others (Mastretta-Yanes et al., 2015) recommend including technical replicates and exploring a range of *de novo* assembly parameter values and their effect on the genotype consistency between different datasets. The optimum parameter profile will be the one that both increases the number of output loci while reducing genotype inconsistencies between replicate samples. By following this approach, Mastretta-Yanes et al. (2015) detected a significant variance in the amount of loci and reliability of the SNP dataset obtained for the non-model plant species *Berberis alpina.* The mean proportion of SNP mismatches between replicate pairs ranged from 2.4% (optimized *de novo* assembly) to 4.2% (*de novo* assembly under default parameter values). These relatively high error rates confirm the importance of evaluating assembly strategies beforehand and understanding the tolerable error rates for the biological question under study. Moreover, an empirical study assessed the effect of bioinformatics pipelines on different population parameters for the Galápagos sea lion (*Zalophus wollebaeki*) (Shafer et al., 2016). They showed that while the variant output (i.e., SNP genotypes) from some pipelines supported outbreeding and population expansion, others, in contrast, supported population bottleneck and inbreeding. Given that the assembly parameters that truly reconstruct genetic variation are ignored, ultimately, appropriate *de novo* assembly will depend on several factors that reflect the genome of the species under study and the experimental goals (Catchen et al., 2011, Catchen et al., 2013).

Bivalves are an ecologically and economically relevant group of marine species. They are among the most polymorphic group of metazoans studied to date (Harrang et al., 2013, Zhang et al., 2012, Saavedra and Bachère, 2006). Hence, *de novo* assembly of bivalve RAD-Seq data is expected to be challenging. This because to reconstruct these highly heterozygous regions a high mismatch value between sequencing reads has to be allowed, increasing the possibility of over-merging non-homologous regions, thereby introducing assembly artefacts (Pryszcz and Gabaldón, 2016). This Chapter aims to establish a *de novo* assembly pipeline of RAD-Seq data from bivalve species. Technical replicates were sequenced from a set of mussel and oyster samples to evaluate the performance of three pipelines for *de novo* assembly of RAD-Seq loci. The pipelines evaluated included: (i) Stacks (Catchen et al., 2011, Catchen et al., 2013), (ii) PyRAD (Eaton, 2014), and (iii) a pseudo-reference approach where we use a combination of Stacks, BWA (Li and Durbin, 2009), and either GATK (McKenna et al., 2010) or SAMtools (Li, 2009) for variant calling (hereafter BWA+GATK and BWA+SAMtools, respectively). The putative optimal pipeline was the one that showed the highest genotype consistency between replicate samples across different bivalve species. The selected pipeline is utilized in future Chapters of this thesis to produce datasets of high confident SNPs to characterize genome-wide patterns of inheritance in bivalve species.

3.2 Material and Methods

3.2.1 Samples

Genomic DNA samples from three species were used to evaluate and optimise a *de novo* assembly and variant calling method for bivalves: the Blue mussel *Mytilus edulis*, the Chilean mussel *Mytilus chilensis*, and the Pacific oyster *Crassostrea gigas*. The quality of *de novo* assembly was assessed by evaluating the genotype consistency between technical replicates. Technical replicates are defined as samples that share the same template DNA (i.e., same individual), but the library preparation steps are performed separately. In our study, technical replicates were sequenced either in the same or different sequencing lanes of an Illumina Hiseq platform.

In total, a set of nine replicate groups (RGs) were evaluated. A summary of the RG information is shown in Table 1. The number of samples processed by technical RG varied from 2 to 3. That is, the same DNA sample (i.e., same individual) was used twice or thrice for library construction. In addition, we included two technical replicates that consisted of different tissue types from a single male or female Blue mussel individual. From a single male (RG 6), four tissue types were sampled: foot, mantle, gill, and adductor muscle. From the single female mussel (RG 5), five different tissue types were sampled; we sampled the gonad in addition to the same tissue types sampled from the male individual. Also, as a control for *de novo* assembly, a control RG that was composed by two tissue samples (gill and fin) from a non-bivalve species, a single male Nile tilapia (*Oreochromis niloticus*) provided by the Institute of Aquaculture (Stirling, UK), was included.

DNA extractions and library preparation methods were followed according to the protocol outlined in section 2.1.3 of Material and Methods (Chapter 2).

Table 1. Summary of RG information

Replicate group	Species	N° of samples	Sequencing lane
RG 1	*M. edulis*	2	different
RG 2	*M. edulis*	2	different
RG 3	*M. edulis*	2	different
RG 4	*M. edulis*	2	different
RG 5	*M. edulis*	5	same
RG 6	*M. edulis*	4	same
RG 7	*M. chilensis*	3	different
RG 8	*C. gigas*	2	same
RG 9	*C. gigas*	2	same

3.2.2 De novo assembly of RAD loci and SNP calling

Paired-end reads were de-multiplexed and filtered for low read quality, missing restriction site, or ambiguous barcodes using the *process_radtag* module of Stacks version 1.40. After quality filtering, we explored the effect of sequentially changing relevant parameters in three *de novo* assembly and variant calling pipelines, Stacks v1.4 (Catchen et al., 2013, Catchen et al., 2011), PyRAD (Eaton, 2014) and a pseudo reference approach that employed a combination of either BWA+GATK-HC or BWA+SAMtools. To the extent possible, equivalent parameter settings were used for pipeline comparison.

An initial single-end *de novo* assembly at default parameters in Stacks was used to identify spurious SNPs. A frequency histogram of the number of SNPs per nucleotide position shows an enrichment of polymorphisms towards the 3' end of the RAD loci assembled from single-end reads (Figure 2). As these SNPs likely represent sequencing errors (Schirmer et al., 2016), single-end reads were trimmed to 100bp using the fastx_trimmer program from FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/).

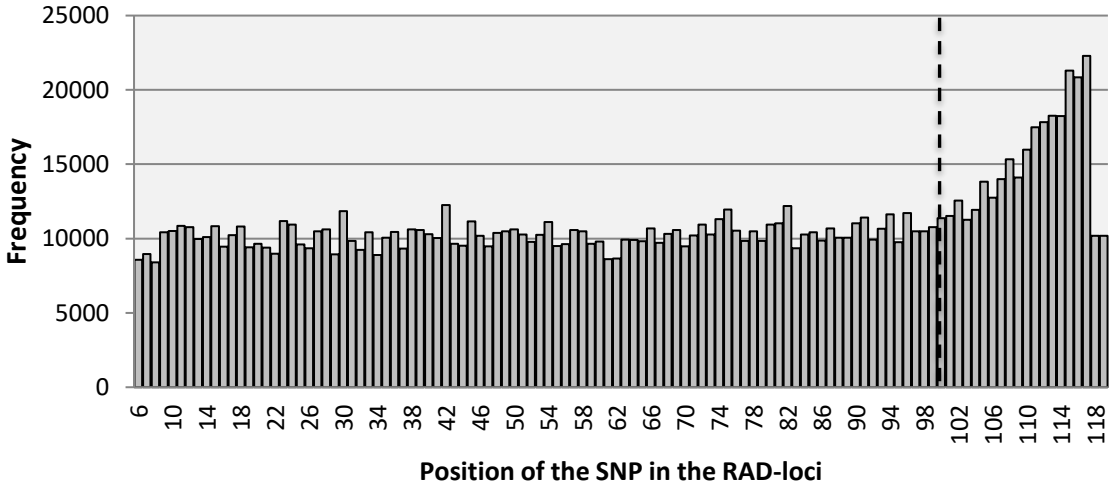


Figure 2. **An increasing number of SNPs is observed after position 100bp on the RAD loci assembled from single-end reads**. The numbers of SNPs below 6 bps are not shown because they represent the unique barcode sequence assigned to each individual.

3.2.3 Stacks

The most widespread approach for processing RAD-Seq data is the Stacks pipeline. Stacks is a software pipeline that organizes RAD-Seq data into loci, anchored by the presence of a restriction enzyme cut site (Baird et al., 2008, Etter et al., 2011a, Etter et al., 2011b), to identify and genotype polymorphisms across individuals. In the absence of a reference genome, *de novo* assembly and SNP calling are performed by Stacks in three stages: first, (RAD) loci are assembled within individuals by identifying loci and their constituent alleles (*ustacks*); then a catalogue of loci is synthethised (*cstacks*); and, finally, individual samples are matched back to the catalogue (*sstacks*) for variant discovery and genotyping. For *de novo* assembly optimization in bivalves, RAD-loci were identified within each individual by changing two core parameters in the *ustacks* module, parameters m and M. We evaluated a minimum read depth to call a stack (i.e., an allele) (-m parameter) of 3, 10 and 30; and a maximum sequence mismatch between stacks (-M parameter) of 1, 6, and 12 bps. A wide spectrum for the M parameter was used in an attempt to assemble highly polymorphic loci. The deleveraging and highly repetitive stacks removal algorithms were enabled. A maximum of six stacks per locus was permitted as a means to intentionally over-merge putative paralogues into single loci. This approach allowed the identification of conflicting RAD-loci (i.e., those with > 2 haplotypes) for further analysis of paralogues. The value for the number of mismatches allowed when building a catalogue/list of RAD-loci across samples (-n) was always set to M+2. Note that a catalogue of loci was created independently for each bivalve species. All combinations of m and M parameter values were tested. Each parameter was changed sequentially while keeping the remaining fixed. Parameters that are not mentioned above were kept to default values. A catalogue of SNPs was retrieved for each RG using the *population* program. To call a variant a minimum of one individual with sequencing information was required. The minimum coverage per SNP marker was set to two times the value used for the m parameter.

3.2.4 PyRAD

The PyRAD pipeline differs to Stacks in that it was designed to analyze highly divergent samples, and may therefore be adequate for *de novo* assembly in species with high polymorphism. Instead of using a strict similarity criterion, PyRAD includes an alignment step before applying a clustering algorithm to identify homologous regions. PyRAD has

been extensively used to recover orthologous loci in distant taxa, allowing, for example, the phylogenetic reconstruction of species that diverged over 17 Myr (Cruaud et al., 2014). One of the anticipated advantages of using PyRAD to accurately perform *de novo* assembly and genotyping in bivalves is that it is able to detect larger indels than Stacks (Sovic et al., 2015). We used the PyRAD pipeline to test three clustering thresholds (Wclust), 88%, 94% and 99% (equivalent to 12, 6 and 1bp mismatch differences in Stacks, respectively), at a minimum depth for a cluster (Mindepth) of 3x, 10x and 30x. The minimum number of samples in a final locus (MinCov) was set to 1. The parameter that identifies/removes paralogues (MaxSH) by controlling the maximum proportion of shared polymorphic sites was disabled, given that technical replicates are expected to share all heterozygous sites.

3.2.5 Pseudo-reference approach: BWA+GATK versus BWA+SAMtools

The pseudo-reference approach consists in defining a representative set of loci from a group of samples, mapping all sample reads back to the pseudo-reference and calling variants based on the alignment. Two variant calling software were tested in our pseudo-reference approach: GATK-HC (McKenna et al., 2010) and SAMtools (Li et al., 2009). To define a representative set of loci for each species, an initial *de novo* assembly of single-end reads was performed using Stacks. The chosen parameter values were the same for all three bivalve species and the Nile tilapia control. Within individuals, reads were assembled into RAD-loci with the *ustacks* module using a minimum stack depth of 10, and a maximum number of mismatches between stacks (or putative alleles) of 6. The presence of indel variation was enabled using the '--gapped-alignment' command. The consensus sequence of the RAD loci from each individual were extracted, grouped by species, and collapsed to a representative set of species-specific loci using three clustering thresholds (to equate to the approach used by the Stacks and PyRAD). The consensus sequences were clustered using a similarity of 88%, 94% and 99% (equivalent to a 12, 6 and 1bp mismatch differences between alleles in Stacks, respectively) with CD-hit (Li and Godzik, 2006, Fu et al., 2012). Single-end reads from all samples were aligned to their respective species pseudo-reference (Blue mussel, Pacific oyster or Chilean mussel reference) using BWA (Li and Durbin, 2009). Reads were aligned without seeding (-l was set to a size larger than read length), and up to two gap openings (-o 2) of up to 12bp (-e 12 -d 12) were allowed. To reproduce the parameter values used in the Stacks and PyRAD pipelines, three values were

evaluated for the maximum number of mismatches allowed between the reference and the sequencing reads: 1, 6 and 12 (equivalent to -n=~0.8, 0.01 and, 0.00001, respectively). Ambiguously mapped reads were removed from the dataset.

SNP calling on the mapped reads was performed using SAMtools and GATK-HC (HaplotypeCaller). Variant discovery was conducted separately for each species. Sample names and read group information were added to the input BAM files using bamaddrg (https://github.com/ekg/bamaddrg). For variant calling in SAMtools, and to avoid spurious mapping around indels, the generated BAM files were subjected to a local realignment with GATK (McKenna et al., 2010). Being a haplotype-based variant detector, GATK-HC does not require a re-alignment step. An initial set of variants were called with the MPILEUP command in SAMtools, with a filter for a minimum mapping score of 20 and a minimum read base quality of 10 (i.e., equivalent to 90% base call accuracy), to match default values in GATK. For variant calling with GATK-HC, default values were used. No hard filters were applied to the GATK dataset to allow for a direct comparison with SAMtools. Markers with missing data for more than 30% of individuals from a RG were excluded. To avoid spurious loci that may have been generated by paralogue (over) merging into a single locus (which is possible at the high mismatch values evaluated), loci containing excess haplotypes were removed (Willis et al., 2017). The species-specific VCF files from the BWA+GATK and BWA+SAMtools pipelines were scanned for common genetic variants using BCFtools (Danecek et al., 2011). To the final dataset of common SNP markers (i.e., same positions at which both the BWA+GATK and BWA+SAMtools pipelines detected a polymorphism), three filters for locus coverage were applied: 3x, 10x and 30x (for comparison with other pipelines).

3.2.6 Preliminary analysis: detection of paralogues in Stacks

In the *de novo* assembly optimization procedure described above, we evaluate different values for the parameter that controls the number of mismatches allowed between alleles. One of the values tested allowed for an extremely high mismatch (parameter M in Stacks equivalent to 12). Thus, we are accepting the possibility that reads that map to a single locus may come from independent loci. This is expected to occur, for example, in cases of undetected paralogy, repetitive elements, and copy number variation. As these conflictive regions are likely to lead to spurious genotype calls, they are typically removed from the

analysis. However, and given our aim is also to get an insight into the genome complexities of bivalves, we evaluated the presence of paralogues in the dataset obtained from the Stacks pipeline. We used the Stacks output because it was obtained without applying stringent filtering for paralogues. For the Stacks pipeline, once RAD loci were assembled within each replicate sample, we extracted (i) the read coverage of alleles at a locus, and (ii) the number of haplotypes and SNP frequency per locus. Evidence of over merging of paralogues was considered if SNP frequencies and haplotype number correlated with coverage.

3.2.7 De novo pipeline comparison

Polymorphic sites produced from the Stacks, PyRAD and the pseudo-reference approach for all parameter combinations were converted into VCF files. The degree of consistency between RGs across both pipelines was assessed by estimating two metrics, following (Mastretta-Yanes et al., 2015): (i) **the Locus error rate**, estimated as the proportion of RAD loci found only in one of the samples of a technical replicate group (i.e., quantifies missing data), and (ii) **the SNP error rate**, measured as the proportion of SNP genotype mismatches (i.e., quantifies genotype discordance) between replicates, without taking into account missing genotypes. Loci (or RAD-loci) are defined as the consensus sequence generated from the assembly of single-end reads. Both metrics were estimated based on the pairwise comparison of replicate samples. It is worth noting that each locus, which in our data extends along 100bp up or downstream of a *SbfI* recognition site, may contain several SNP markers.

The best pipeline for *de novo* assembly and SNP calling for bivalves was considered to be the profile of parameter combinations that minimized locus and SNP error rates, irrespective of the number of RAD loci or SNPs identified. Although some authors (Mastretta-Yanes et al., 2015) suggest choosing the set of parameters that minimize errors and maximize the retrieval of informative loci, our analysis revealed that by aiming at maximizing the number of (RAD) loci we increasingly identified highly polymorphic regions. These regions contain information from contiguous SNPs and will mostly add redundancy instead of reflecting an independent genetic signal.

3.3 Results and Discussion

To develop and optimise a pipeline for *de novo* assembly and SNP calling of bivalve RAD-Seq data we evaluated the genotype consistency of 9 RGs that underwent three different assembly procedures – performed either by Stacks, PyRAD or a pseudo-reference approach.

After quality filtering, the number of reads across samples averaged 2,399,400. The mean GC content of the retained reads was 38% for the Pacific oyster, 36% for the British mussel, and 36% for the Chilean mussel. Two samples that belonged to RGs 7 and 8, which were composed by four and three technical replicates each, were removed from the analysis due to their low number of quality reads (< 500,000). The number of reads between replicate samples varied significantly, with some samples showing more than twice the standard deviation from the mean number of reads. The fact that we also observed significant read variation within technical RGs that were sequenced in the same sequencing lane suggests that significant technical noise accumulates during the library preparation process. To reduce the bias due to sequencing coverage variation across samples, sequences were normalized randomly to the sample with the lowest post-quality filtered read count per species – 600,000 for the Pacific oyster, 1,500,000 for the Chilean mussel, and 770,000 for the British mussel.

3.3.1 Preliminary analysis: detection of paralogues in Stacks

Prior to the comparison of the different pipelines tested, we assessed the presence of putative paralogues in the bivalve RAD-Seq data assembled with Stacks. As expected, the distribution of the average number of SNPs per RAD loci varied with different combinations of *de novo* assembly parameters. However, the mode of the SNP frequency distribution was extreme at the highest value tested for the parameter that controls the number of mismatches allowed between haplotypes (M =12 in Stacks) – ~3 SNPs per RAD loci for both the British mussel and Pacific oyster samples (Figure 3). This large SNP frequency was consistently observed at all the coverages examined (locus read-depths = 3x, 10x and 30x). Therefore, before the pipeline comparison, we evaluated the validity of these RAD loci, as they may represent spurious loci derived from the 'over-merging' of paralogues.

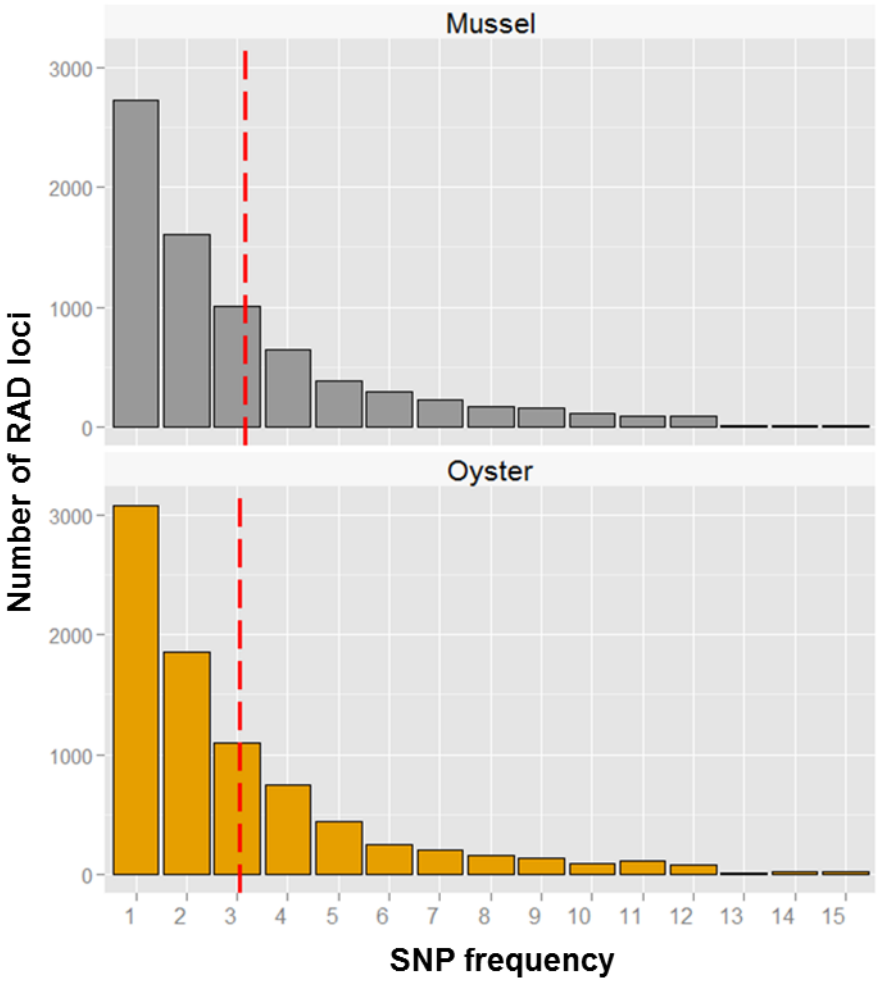


Figure 3. A high SNP frequency is observed across British mussel and Pacific oyster RGs when RAD loci are assembled at a mismatch value of 12bp. The mean SNP frequency per RG is indicated by a red dotted line.

We assessed the possibility that this dataset harboured a high SNP frequency due to the over-merging of paralogues by (i) evaluating the correlation between SNP frequency and coverage and (ii) counting the number of haplotypes per assembled locus. Under the scenario that a high SNP frequency was caused by the over-merging of non-homologous genomic regions, a positive correlation between coverage and SNP frequency may be expected. The correlations between SNP frequency and coverage was low but significant for all RGs; Pearson's correlation coefficient averaged -0.18 for the British mussel (p-value<2.2E-16), -0.015 for the Chilean mussel (p-value=0.003), and -0.028 (p-value=8.5E-12) for the Pacific oyster. The second metric we used to detect the potential merging of paralogues into single loci was to count the number of haplotypes observed for all the RAD

loci assembled within individuals. Bivalves are diploid (Thiriot-Quievreux, 2002). Thus the maximum number of haplotypes expected for any locus representing a single genomic region is expected to be two. A histogram of haplotype number per locus shows a greater frequency of bi-allelic RAD loci (i.e., two haplotypes) at all M parameter values tested in Stacks (1, 6 and 12). A significant increase in the proportion of bi-allelic RAD loci was observed for all RGs when a higher mismatch value was used for *de novo* assembly (test for equality of proportions between M=1 and M=12; p-value<0.0001 for the Blue mussel, Chilean mussel and Pacific oyster RGs). Further evaluation revealed that this increase is due to the merging of monomorphic RAD loci into regions of high nucleotide diversity, as a change in the M parameter value from 1 to 12 increases the SNP frequency of the polymorphic loci from 1.06 to 2.97 for the Chilean mussel, 1.22 to 3.01 for the Pacific oyster, and 1.14 to 3.12 for Blue mussel. These high diversity regions appear to be not associated with the over-merging of paralogue regions, as no significant increase in the proportion of loci showing 3, 4 or >4 haplotypes was observed after allowing for a higher nucleotide mismatch between alleles (Figure 4) (test for equality of proportions between M=6 and M=12; p-values for all haplotype categories within RGs was >0.11). In the same line of evidence, similar profiles of histograms of coverage are observed for loci showing different SNP frequencies (Figure 5). Taken together this evidence suggests that bivalve genomes harbor high diversity regions, as when we allow for the reconstruction of highly polymorphic regions (i.e., at high M parameter values), no significant increase in average coverage per locus is observed. The opposite would be expected under the assumption that these high diversity regions emerge due to the 'over-merging' of highly similar but loci (i.e., paralogues), as this would lead to an increase in the average coverage of the locus.

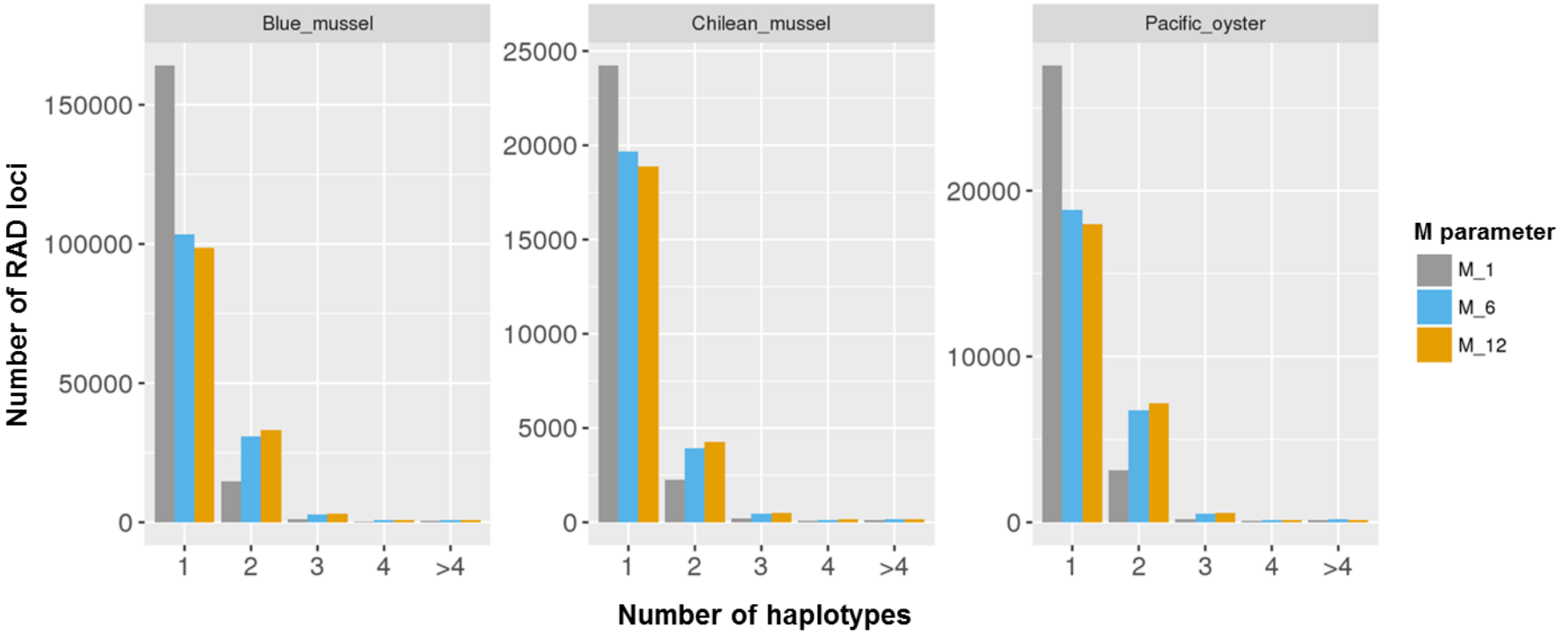


Figure 4. A higher number of monomorphic (1 haplotype) and polymorphic (2 haplotypes) RAD loci are observed across different mismatch values tested (1, 6, and 12) compared to spurious loci (i.e., those with >2 haplotypes).

,

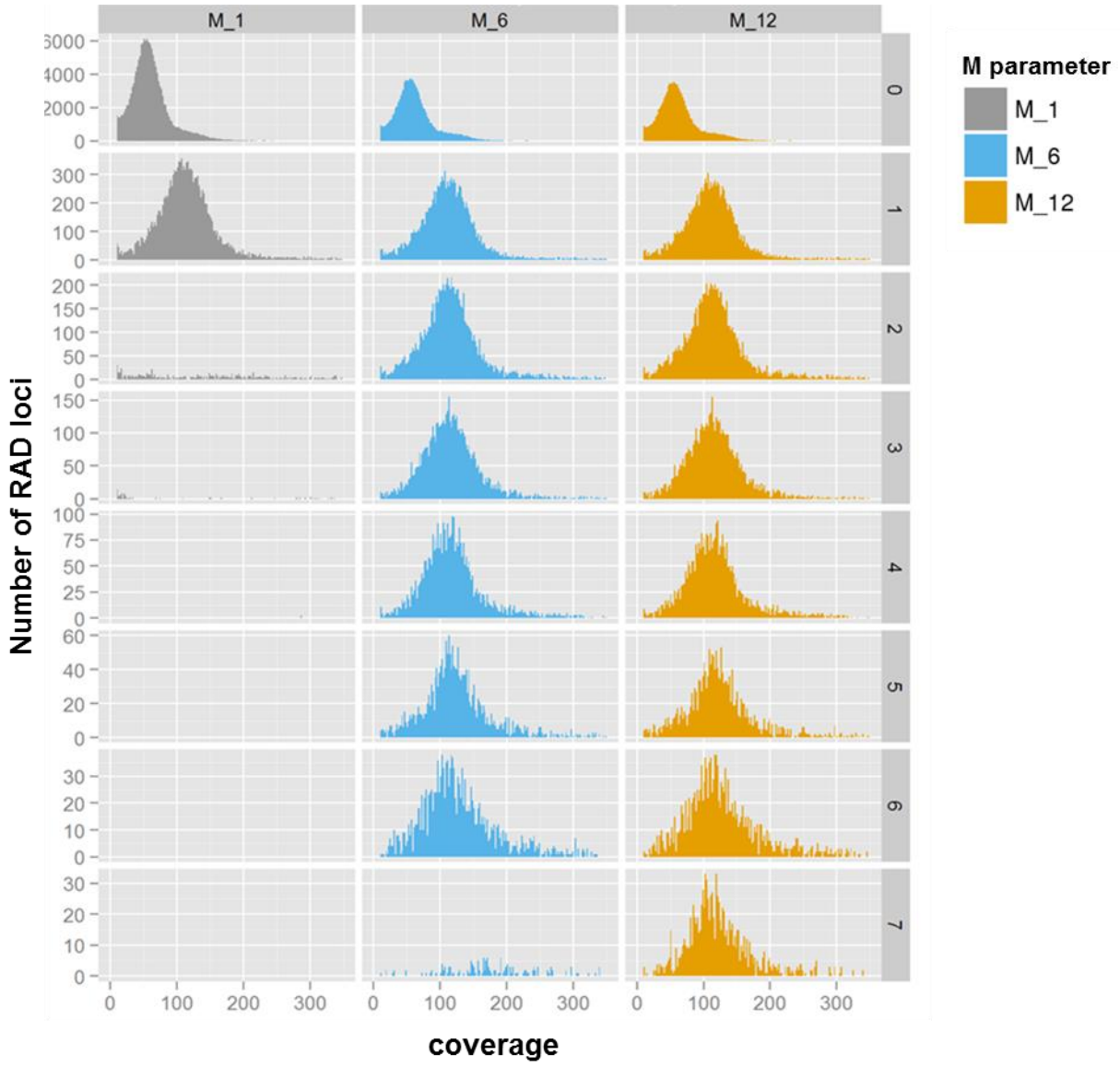


Figure 5. Evaluation of the effect of varying different M parameters (1, 6 and 12) on the coverage of RAD loci with different SNP frequencies (0, 1, 2, 3, 4, 5, 6 and 7 SNPs per RAD loci). Coverage is shown on the x-axes. In the right y-axes the SNP frequencies observed in the RAD-loci. In the left y-axes the frequency of RAD loci. Each column (color coded) represents the different M parameters tested and the distribution obtained for RAD loci with different SNP frequencies (stratified by rows). Note that the values on the y-axes (Number of RAD loci) were allowed to vary independently for each SNP frequency category.

Although we do not find a significant association between coverage and polymorphism rate of the assembled loci, which may have been indicative of 'over-merging,' we find a strong association between coverage and the polymorphic status of a locus (i.e., whether a locus is categorised as monomorphic or polymorphic). A monomorphic locus is defined when the two alleles of diploid individuals are identical. In contrast, a polymorphic locus is defined when there is sequence variation between two alleles at a locus. After the *de novo* assembly of RAD-loci within individuals, we observe a bimodal distribution of the coverage (per locus) histogram. Interestingly, the two modes of the distribution correspond to the peak coverage of the monomorphic and polymorphic RAD-loci. These two coverage peaks of unknown copy number are separated by approximately twice the median coverage. This pattern is strikingly consistent, as it was observed for all the values of allele mismatch (M) and coverage (m) tested in Stacks (Figure 6). Moreover, given that the control RG (i.e., the Nile tilapia replicate) shows a unimodal distribution for both monomorphic and polymorphic categories (Figure 6), we conclude that the two peaks of loci present in the mussel and oyster species likely reflect a genome property of bivalves.

This type of distribution has been observed in species that experienced recent genome duplication events, such as salmonid fishes (McKinney et al., 2016), and have been attributed to spurious assemblies generated by paralogue over-merging. Although a plausible explanation for salmonids, it would be difficult to explain why a high proportion of polymorphic RAD-loci are a by-product of paralogue over-merging (referred to hereafter as hypothesis 1). An alternative, reciprocal explanation would be that the true diploid state is represented by the higher coverage peak and that the monomorphic mode represents hemizygous loci, which would explain why the depth of coverage is exactly half for the monomorphic RAD-loci (referred to hereafter as hypothesis 2). A hemizygous locus is that for which one allele is missing, as a result, for example, of intragenic deletions. Under this scenario, however, it would be difficult to explain why most of the monomorphic loci lack genetic information for the other allele at a locus, whether the missing allele is identical (true monomorphic) or not (true polymorphic). Despite evidence supporting aneuploidy as a common phenomenon in bivalves (Rico et al., 2017, Thiriot-Quiévreux et al., 1992), the levels would have to be exceptionally high to explain the unusual distribution observed in Figure 6.

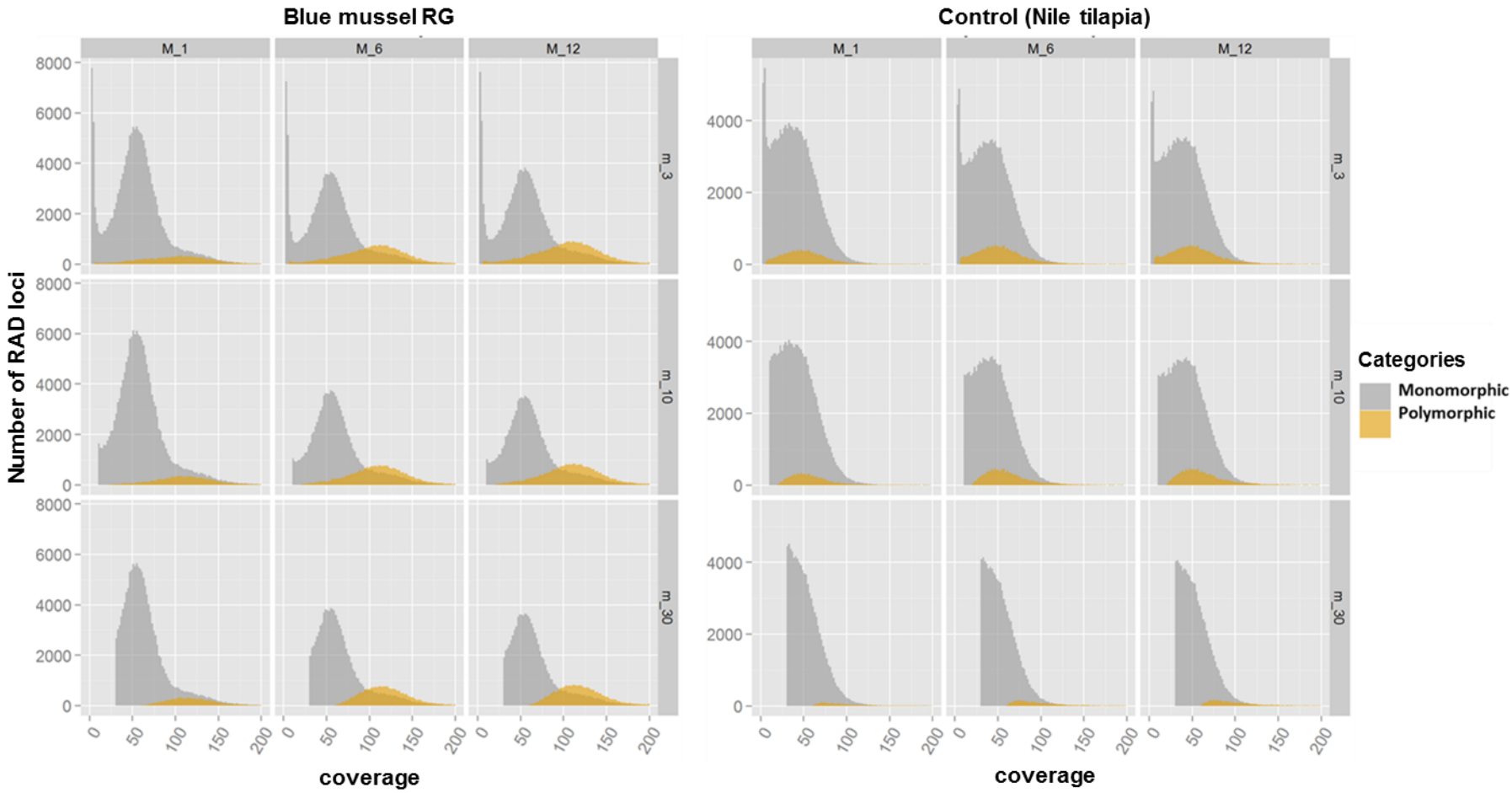


Figure 6. Histogram of RAD-loci coverage at different combinations of M and m Stacks assembly parameters values for the (left) Blue mussel RG and (right) a Nile tilapia control.

We further examined the hypothesis that the average coverage of polymorphic loci is higher than their monomorphic counterparts due to the merging of paralogues (hypothesis 1). The prediction is that if reads from distinct but highly similar loci are being merged into a single locus, then this should be reflected in the ratio of the alleles. To fit the distribution that we observe in Figure 6, we would have to be 'over-merging' two duplicated regions (to explain the 2-fold coverage of polymorphic over monomorphic loci). Under the scenario of assembly artefacts caused by the over-merging of two loci, it would be more likely to observe a bias towards 1:3 allelic ratio compared to a 1:1 ratio, as it is more probable that the difference between two paralogue regions is a single bp (e.g., C/A + A/A = 1:3 ratio) rather than two bp (C/C + A/A= 1:1 ratio), which would require two consecutive mutational steps. The allele dosage of heterozygous RAD loci with less than 500x coverage was estimated. As shown in Table 2, most of the allele ratios of di allelic RAD loci conform to the 1:1 ratio predictable for heterozygote Mendelian loci. Any deviations from Mendelian expectations may be caused by technical variations derived, for example, from an extended PCR step (Heinrich et al., 2012). We also tested whether the reads of a locus conformed to a 1:3 ratio, given that conformity to this ratio would be indicative of paralogue over-merging. A significant amount of the allelic fractions of the RAD loci (44% to 46%) followed a predicted 1:3 ratio, after adjusting for multiple testing in a chi^2 test. This evidence would indicate that the aforementioned polymorphic RAD loci are not necessarily Mendelian loci, and could, in fact, be the result of paralogue over-merging. In general, the evaluation of allele ratios suggests that most RAD-loci are single Mendelian loci, as they tend to follow a 1:1 allele ratio, although a fraction appears to follow the allelic ratio pattern of paralogue over-merging.

We have no means at this stage of our analysis to evaluate whether the bimodal distribution of coverage is caused by extensive hemizygosity in the bivalve genome (hypothesis 2).

Table 2. Number of SNPs across all bivalve RGs that follow a 1:1 and a 1:3 allele ratio at different M parameter values. In parenthesis the percentage of SNPs tested that conformed to the expectation of the ratio between alleles.

		Allele mismatch (parameter M)		
Genotype	**Allelic ratio**	**1**	**6**	**12**
AABB	1:1	60,491 SNPs (83%)	128,500 SNPs (88%)	140,676 SNPs (88%)
ABBB or AAAB	1:3	32,190 SNPs (44%)	65,975 SNPs (45%)	72,625 SNPs (46%)

Overall, the evidence presented herein indicates that irrespective of the *de novo* assembly parameters that we use, two conspicuous groups of RAD-loci are detected in the bivalve genome (example presented in Figure 6). One group shows a comparatively moderate coverage (~50x) and mostly comprises monomorphic loci, whereas the other group is characterized by a higher coverage (~100x) in addition to a higher proportion of polymorphic loci. This double coverage-depth relationship between peaks has been observed previously in whole genome sequencing projects of bivalves (Zhang et al., 2012, Takeuchi et al., 2012), and has been interpreted as an artifact caused by genome heterozygosity. Because of a high divergence between alleles, reads from one allele can only map onto the haploid (allele) sequence they came from but cannot map to the homologous locus, leading to a group of loci with half-coverage. A further analyses attempting at deciphering whether this abnormal distribution is generated by null alleles is explored in Chapter 5.

3.3.2 Comparison of *de novo* assembly pipelines for bivalve RAD-Seq data

Bivalve species have a highly polymorphic genome, with an average density of SNPs being equally high in both coding and non-coding regions (one SNP every 60bp vs. one every 40bp, respectively; Sauvage et al. (2007)). De novo assembly of polymorphic genomes is challenging, mainly because of the difficulty of handling highly divergent haplotypes, and appropriately reconstructing highly polymorphic loci in a single individual. Hence, to obtain a reliable set of genetic markers of bivalve species, it is critical first to optimize *de novo* sequence assembly parameters. We compared three different pipelines for RAD-Seq data processing, using both technical and biological replicates to assess the robustness of these variant calling methods. Two of the evaluated pipelines were Stacks and PyRAD. In addition, we explored a 'pseudo-reference' approach in which a pseudo-reference genome was created, reads were mapped with BWA, variants were called either with GATK-HC or SAMtools, and the intersection of these two variant calling files was used as a high confident SNP dataset (Figure 7).

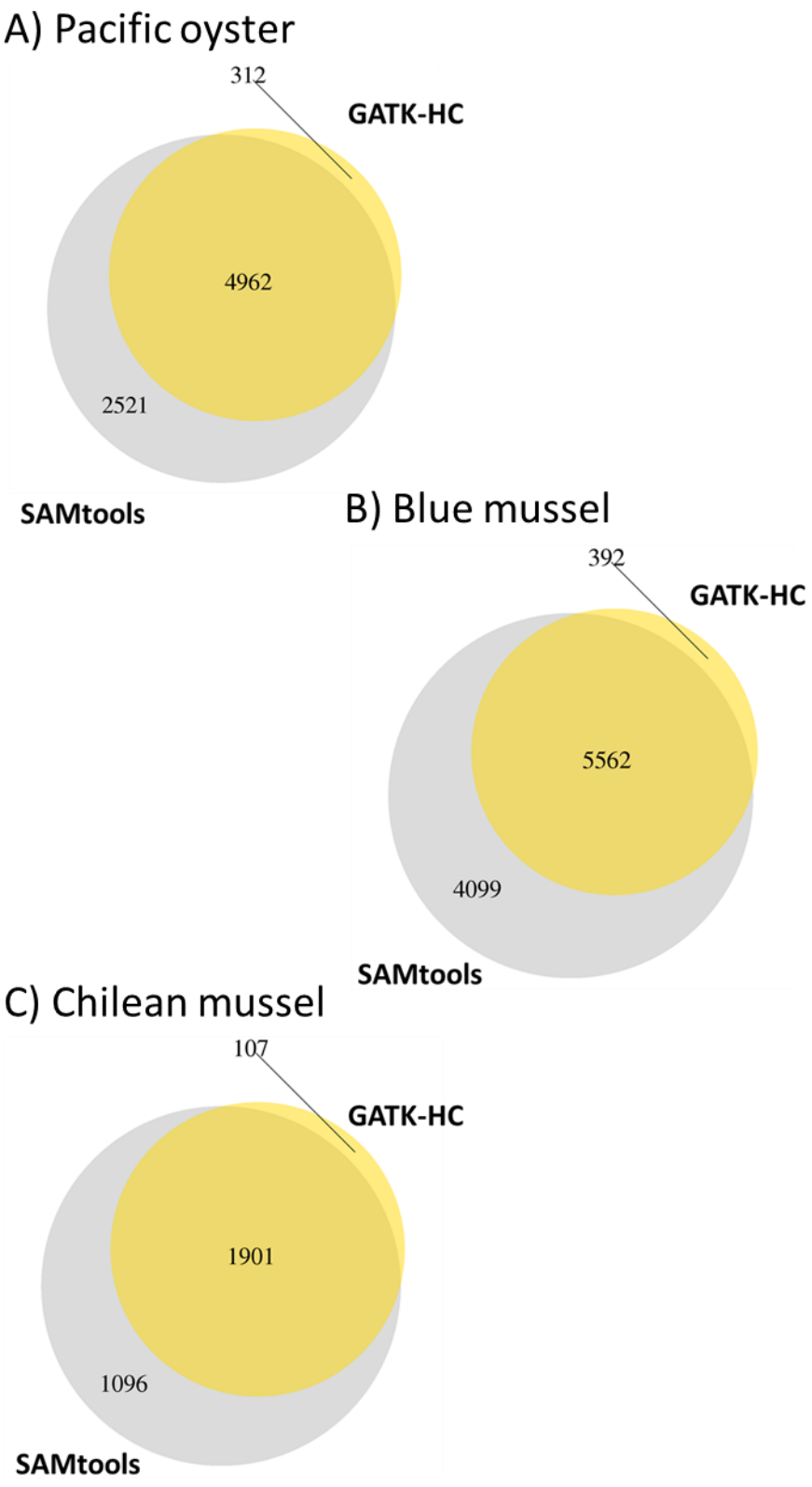


Figure 7. Number of common SNPs identified by two variant calling pipelines, GATK-HC and SAMtools, across the RGs from three different bivalve species.

A total of 9 RGs (Table 1) were assembled *de novo* and used for SNP calling via three different pipelines. To facilitate the comparison, RGs were combined by species: Pacific oyster (*C. gigas*), Blue mussel (*M. edulis*), or Chilean mussel (*M. chilensis*). Results obtained from the three pipelines show intra and inter-pipeline variation regarding (i) number of RAD-loci and SNPs identified, (ii) SNP error rate, defined as a genotype disagreement between replicate pairs, and (iii) Locus error rate, defined as instances in which a sample from a replicate pair has missing data whereas the other has been genotyped.

3.3.3 Pipeline comparison: Number of RAD loci and SNPs

Regarding the total amount of RAD-loci and SNPs identified, we noted that the Stacks and PyRAD pipelines numbers varied considerably, although consistently, across the two major *de novo* assembly parameters tested for optimization, m and M. These parameters represent the minimum coverage to call an allele (m) and the maximum number of nucleotide differences allowed between alleles (M). For both pipelines, the effect of the coverage was to reduce the number of RAD-loci identified, as expected (example in Figure 8). For instance, increasing the minimal allele coverage from 3x to 30x led to a 30% and 50% reduction in the number of RAD-loci identified in Stacks and PyRAD, respectively. This reduction in the number of RAD-loci identified at higher coverages also affects the number of SNPs discovered; at higher coverages the frequency of variants is reduced due to the lower amount of regions available for variant screening.

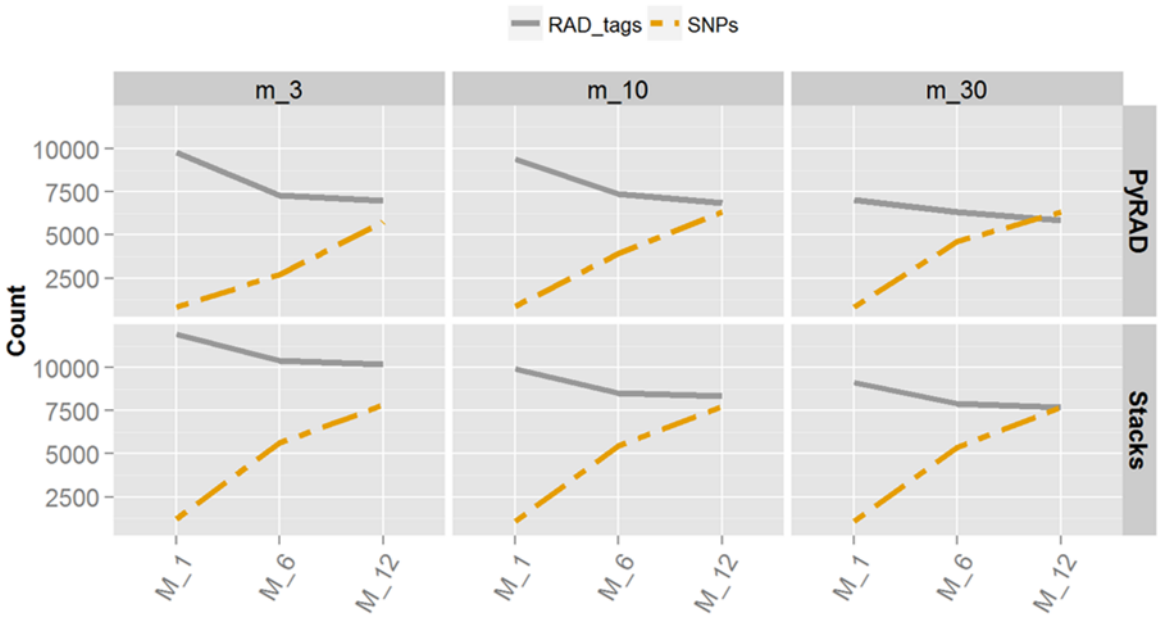


Figure 8. Representative example of the relationship between the (i) total number of RAD-loci (solid grey line) loci and (ii) SNPs (dotted yellow line) identified after changing the threshold of allele coverage (m) and the number of mismatches allowed between alleles (M) in the Stacks and PyRAD pipelines.

The most relevant *de novo* assembly parameters controlling the number of SNPs identified is parameter M. Indeed, we found that the maximum number of mismatches allowed between alleles (parameter M and WClust, in Stacks and PyRAD, respectively) is a key source of variation for bivalve RAD-Seq data. Because bivalves exhibit extreme levels of sequence polymorphism, allowing for more divergence between alleles (Figure 3) enabled the recovery of putatively single loci harboring high SNP frequencies. For polymorphic genomes, this parameter is expected to play a major role when reconstructing true patterns of biological variation. If this threshold parameter is set too low, a locus that contains dissimilar alleles (i.e., highly polymorphic regions) will be split into two different monomorphic loci (over-splitting). Inversely, if this parameter is set too high, there is a risk of assembling paralogue regions into single loci (under-splitting). Of the three M parameter values tested (1, 6 and 12), the total number of SNPs increases dramatically across mismatch vales at the time monomorphic loci decrease (Figure 8), which suggests that at relaxed values (M=12) monomorphic loci are allowed to merge into high diversity loci. Nevertheless, there is a possibility that a fraction of these regions resulted from the

merging of paralogue containing regions, as no filter for paralogy was applied in the Stacks and PyRAD pipelines.

A direct comparison of the number of RAD-loci and SNPs with the 'pseudo-reference' approach was not possible. Due to the fact that variants from this approach were filtered for coverage at the end of the analysis, SNPs that did not pass the threshold were set to missing by the software instead of being removed from the dataset, making it difficult to quantify an overall effect for comparison.

3.3.4 Pipeline comparison: SNP and Locus error rates

As mentioned above, SNP error rates and Locus error rates show intra and inter-pipeline variation. SNP error rates are defined as the fraction of SNPs typed in a replicate pair for which genotypes between samples disagrees, after the removal of markers with missing data. Hence, a low SNP error is an indication of higher accuracy and better assembly parameter values. A Locus error rate is defined as the fraction of SNPs typed in a replicate pair for which one sample is successfully genotyped and the other exhibits a missing genotype. First, results are discussed for each pipeline separately, followed by a cross-comparison of pipelines for the same species-specific replicates.

Results for the Stacks pipeline show that the SNP error rate is maintained relatively constant across replicates from the same species for both the m and M parameters, except for the M parameter in the Pacific oyster set of replicates (Figure 9A). A comparison across species indicates that, overall, the error is significantly higher for the Chilean mussel replicates, as it ranged from 20% to 40%. The reason for this high error rate may have been a bias caused by the lower DNA quality of this set of replicates or because technical replicates from this group were sequenced in different platforms (Hiseq2500 vs. Hiseq2000). For the Blue mussel replicates, errors ranged from 0.6% to 6%, whereas for the Pacific oysters replicates the SNP error rate ranged from 2% to 13%. Regarding the Locus error rate, a higher variation across parameters values and species was observed (Figure 9B). At the species level, Chilean mussel replicates show higher Locus error rates at most of the parameter values tested, ranging from 9% to 19%. This higher error rate, as mentioned above, may have been associated to the fact that technical replicates for this RG were sequenced in different platforms. However, because we did not formally evaluate inter-platform variability, this hypothesis cannot be tested. When analyzing the m parameter

(i.e., the minimum number of reads to call an allele) values, at the highest value tested (m=30), an overall increase in Locus error rates is observed in both the Blue mussel and Pacific oyster replicates, suggesting the occurrence of allelic dropout. Oppositely, Locus error rates are higher at the lowest m parameter values tested (m=3) for the Chilean mussel replicates.

For the PyRAD pipeline, SNP error rates were higher for the Chilean mussel replicates, consistent with the results from Stacks (Figure 10A). However, SNP error rates were found to be higher compared to Stacks, ranging from 16% to 40%. The fact that these error rates were greater than those obtained from the Stacks pipeline suggests that PyRAD is misplacing sequencing reads, inflating genotyping errors. The SNP error rates for the Blue mussel replicates ranged from <1% to 9%. On the other hand, the SNP error rates for the Pacific oyster replicates ranged from 3% to 9%. The effect of the different m parameter values tested had a marginal effect on SNP error rates for both the blue mussel and the Pacific oyster; whereas for the Chilean mussel, at higher allele read-depths lower SNP error rates were observed, probably indicating that the process of genotype calling in PyRAD performs better at higher depths. No clear difference is observed for SNP error rates across species-specific replicates when the M parameter is changed. Regarding the Locus error rate in PyRAD, no clear pattern across m and M parameters is observed across species-specific replicates (Figure 10B). Nevertheless, a slight increase in Locus error rates is observed at m=30 in both the Blue mussel and Pacific oyster replicates, similar to what was observed in the Stacks pipeline, and probably an indication of allelic dropout.

Finally, for the pseudo-reference pipeline, SNP error rates were lower compared to those obtained for the Stacks and PyRAD pipelines for all species-specific replicates. SNP error rates reached 0.6% for the Blue mussel replicates, 1.9% for the Pacific oyster replicates, and 7% for the Chilean mussel replicates (Figure 11A). No significant difference in the range of the SNP error rates was observed when the parameter that controls for locus depth (m) was set to 3, 10 or 30 in the Blue mussel group of replicates. For the Chilean mussel replicates, the same tendency was observed across the different minimum locus read-depths evaluated (m= 3, 10 or 30) at different M parameters (i.e., mismatch values between alleles). For the Pacific oyster replicates, the range of SNP errors was different depending on the m parameter value used. Particularly higher ranges of SNP error rates were detected in the Pacific oyster replicate group at lower read-depths per locus (m=3) if a

mismatch between alleles of one or six is allowed (0.6%-5%); whereas at the highest locus read-depth tested (m=30) the ranges of SNP error rates were minimal (0.25%-1.5%). On the other hand, regarding the Locus error rate (i.e., estimate that evaluates whether genotypes are present in both replicate samples), the lowest Locus errors are detected in the Chilean mussel replicates, ranging from 0.4% to 5.7%, taking into consideration all combination of m and M parameter values tested (Figure 11B). The highest Locus error rates are detected for the Pacific oyster replicates, ranging from 1% to 15%, considering all combination of m and M parameter values tested. Across m parameter values, the locus read-depth of 30 (m=30) generates the highest Locus error rates across species replicates, suggesting that setting this value too high causes locus dropout, as expected. This trend is observed for species-specific RGs, although again is more severe in oysters.

To simplify the visualization of the optimal pipeline and combination of parameter values for *de novo* assembly of bivalve RAD-Seq data, the SNP error rates obtained for all three pipelines were graphed together for a (i) the Blue mussel replicate group (the species-specific group of replicates that comparatively showed lower error rates) and the (ii) Chilean mussel replicate group (the species-specific group of samples that performed worse, with high error rates). We took the approach of comparing the pipelines in 'good' and 'bad' quality samples to account for instances in which a mixture of DNA qualities and sequencing qualities is retrieved, which is the most common scenario in genetic analysis.

As shown in Figure 12, the lowest error rates for the 'good' quality replicate group (i.e., the Blue mussel *M. edulis* RGs), were identified with the pseudo-reference approach at all combination of m and M parameter values tested, with an average SNP error rate of 0.6%. This implies that good quality RAD-Seq data that is processed with the pseudo-reference pipeline should yield high confidence SNP genotypes in a species with high sequence polymorphism. On the other hand, the worst performing pipeline for the RAD-Seq data was PyRAD, with errors averaging 3.3%. Since all pipelines are processing the same sequencing data, the difference in error rates is reflecting differences in the internal algorithms of the software. The most commonly used pipeline to process RAD-Seq data, Stacks, averaged a SNP error rate of 3% across all combination of m and M parameter values.

When the pipelines were used to assemble 'bad' quality RAD-Seq data (represented by the Chilean mussel RGs), the pseudo-reference also performed better than the PyRAD and Stacks pipelines (Figure 13). SNP error rates averaged 7% for the pseudo-reference

approach, whereas the average was 23% for the PyRAD pipeline and 29% for Stacks. Overall, all pipelines perform poorly for ‘bad’ quality samples, although some more than others.

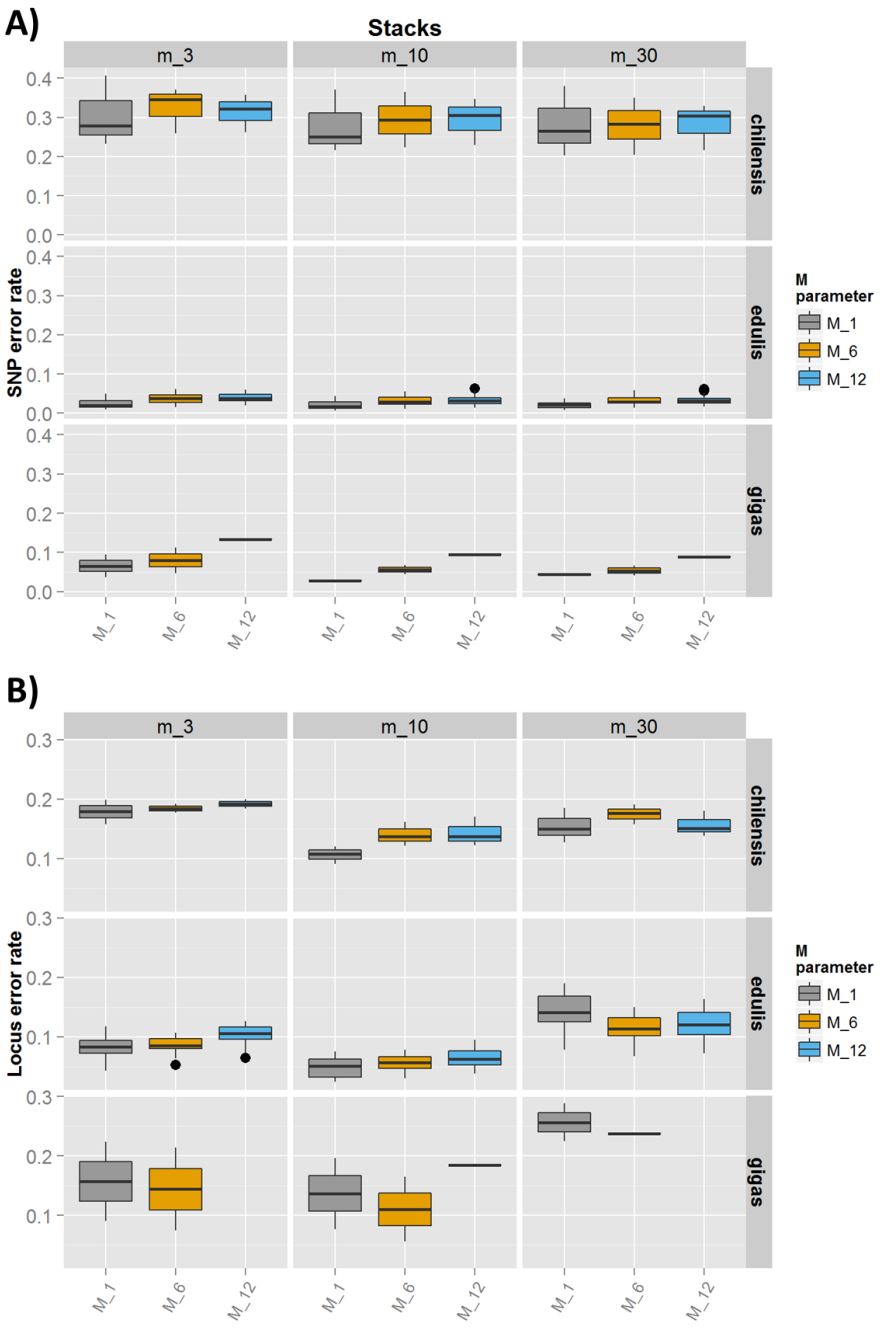


Figure 9. Effect of two different assembly parameter values on the (A) SNP error rate and (B) Locus error rate for RAD loci obtained from the Stacks pipeline in three different species-specific RGs. In the x-axes the different M parameter values tested are shown. In the y-axes the proportion of markers showing either a SNP or a Locus error is shown. Parameter m is the minimum coverage to call an allele and parameter M is the maximum number of nucleotide differences allowed between alleles. Chilensis stands for *M.chilensis* (the Chilean mussel), edulis stands for *M.edulis* (blue mussel), and gigas stands for *C. gigas* (the Pacific oyster). Outliers are represented by black dots.

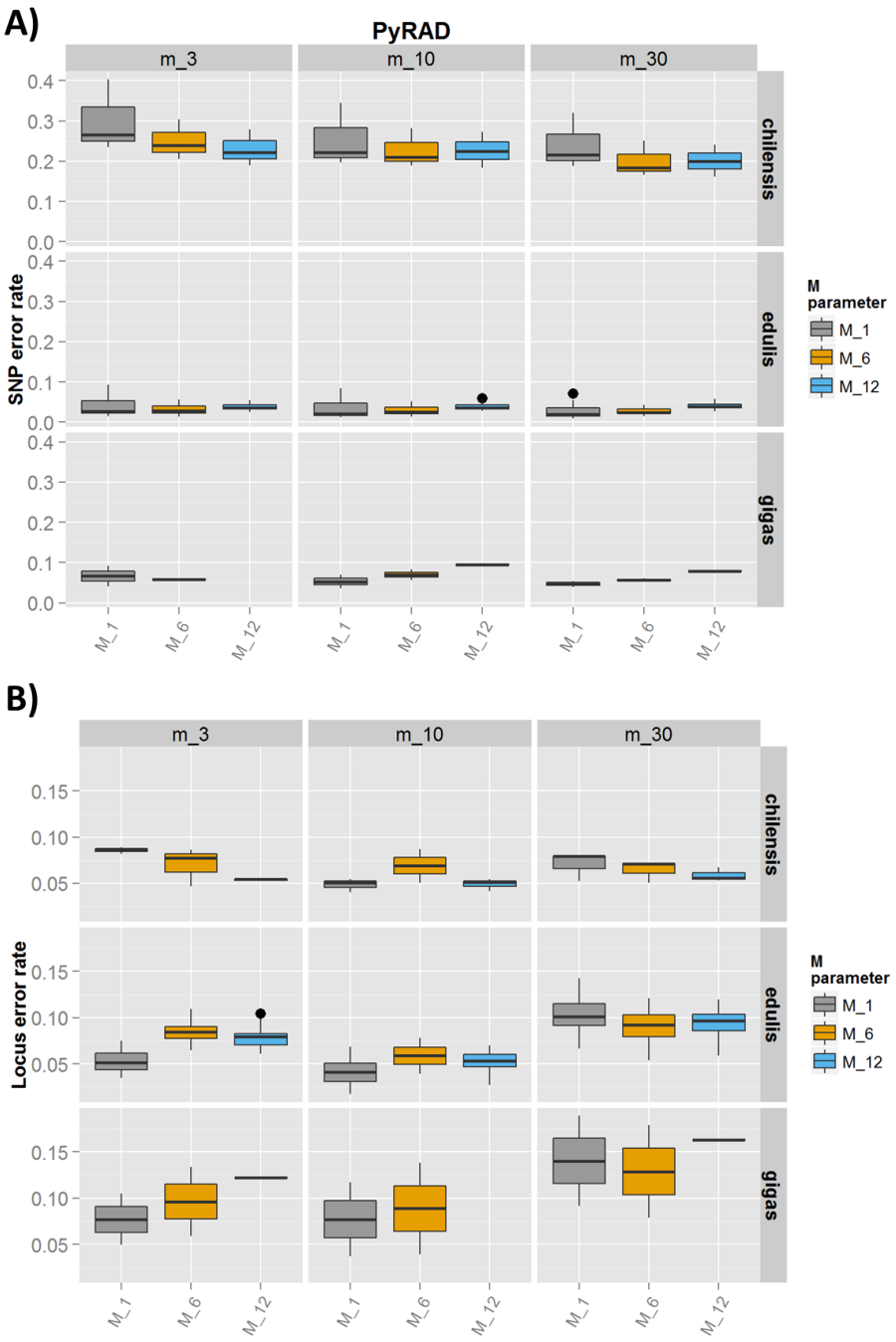


Figure 10. Effect of two different assembly parameter values on the (A) SNP error rate and (B) Locus error rate for RAD loci obtained from the PyRAD pipeline in three different species-specific RGs. In the x-axes the different M parameter values tested. In the y-axes the proportion of markers showing either a SNP or a Locus error. Parameter m is the minimum coverage to call an allele and parameter M is the maximum number of nucleotide differences allowed between alleles. Chilensis stands for *M.chilensis* (the Chilean mussel), edulis stands for *M.edulis* (blue mussel), and gigas stands for *C. gigas* (the Pacific oyster). Outliers are represented by black dots.

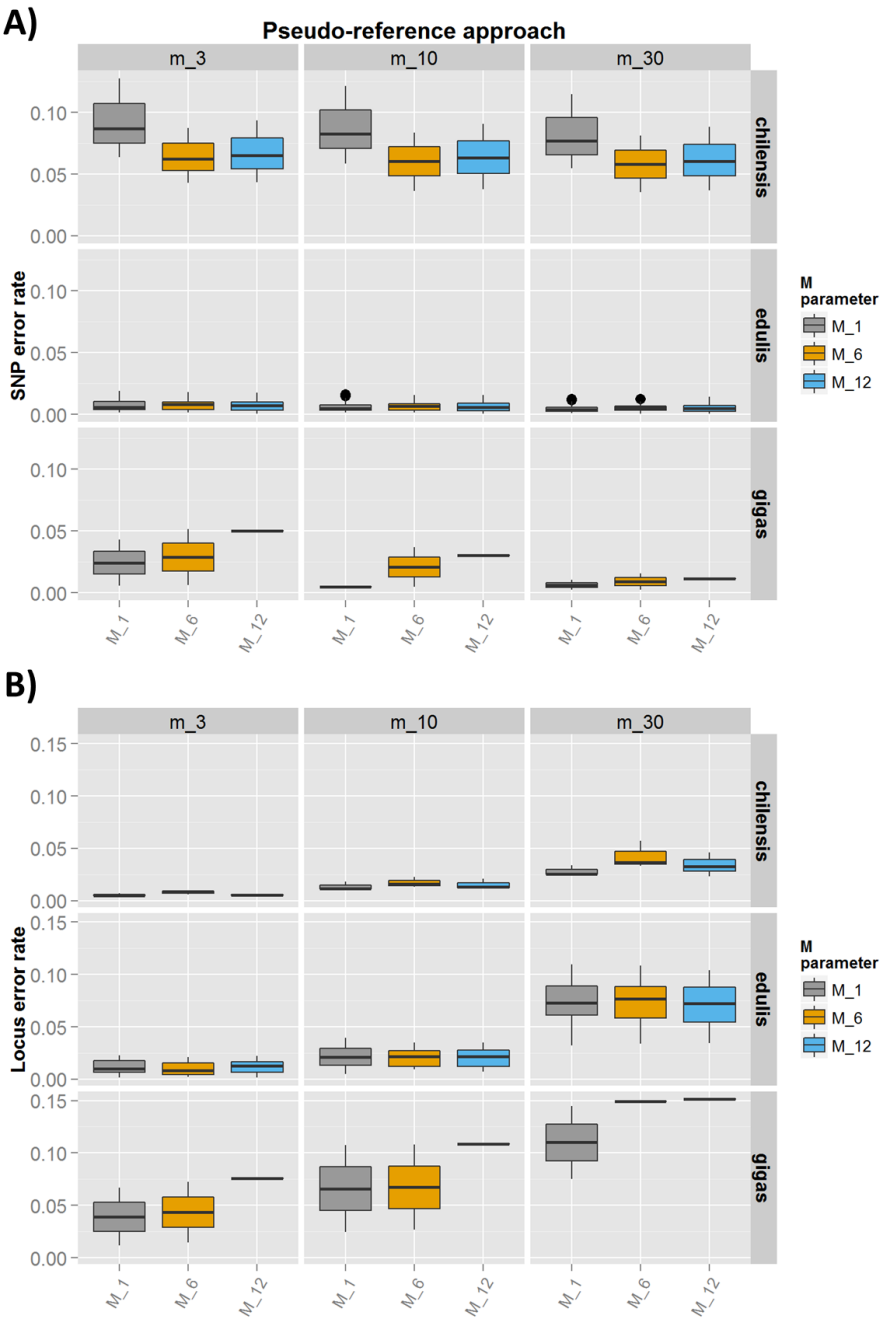


Figure 11. Effect of two different assembly parameter values on the (A) SNP error rate and (B) Locus error rate for RAD loci obtained from the 'pseudo-reference' pipeline in three different species-specific RGs. In the x-axes the different M parameter values tested. In the y-axes the proportion of markers showing either a SNP or a Locus error. Parameter m is the minimum coverage to call an allele and parameter M is the maximum number of nucleotide differences allowed between alleles. Chilensis stands for *M.chilensis* (the Chilean mussel), edulis stands for *M.edulis* (blue mussel), and gigas stands for *C. gigas* (the Pacific oyster). Outliers are represented by black dots.

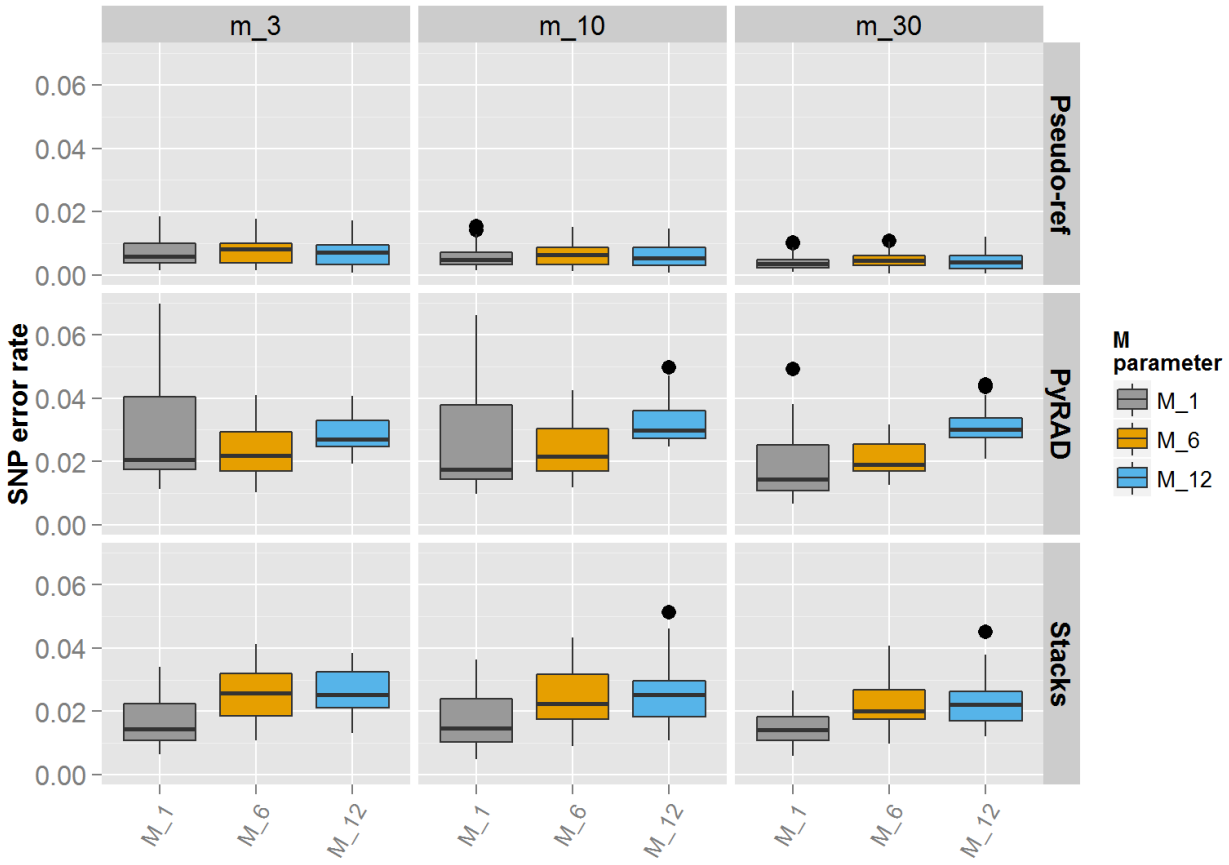


Figure 12. Comparison of the SNP error rates obtained from three different pipelines for a 'good' quality RG at different combinations of m and M parameters. In the x-axes the different M parameter values tested. In the y-axes the proportion of markers showing a SNP error. Parameter m is the minimum coverage to call an allele and parameter M is the maximum number of nucleotide differences allowed between alleles.

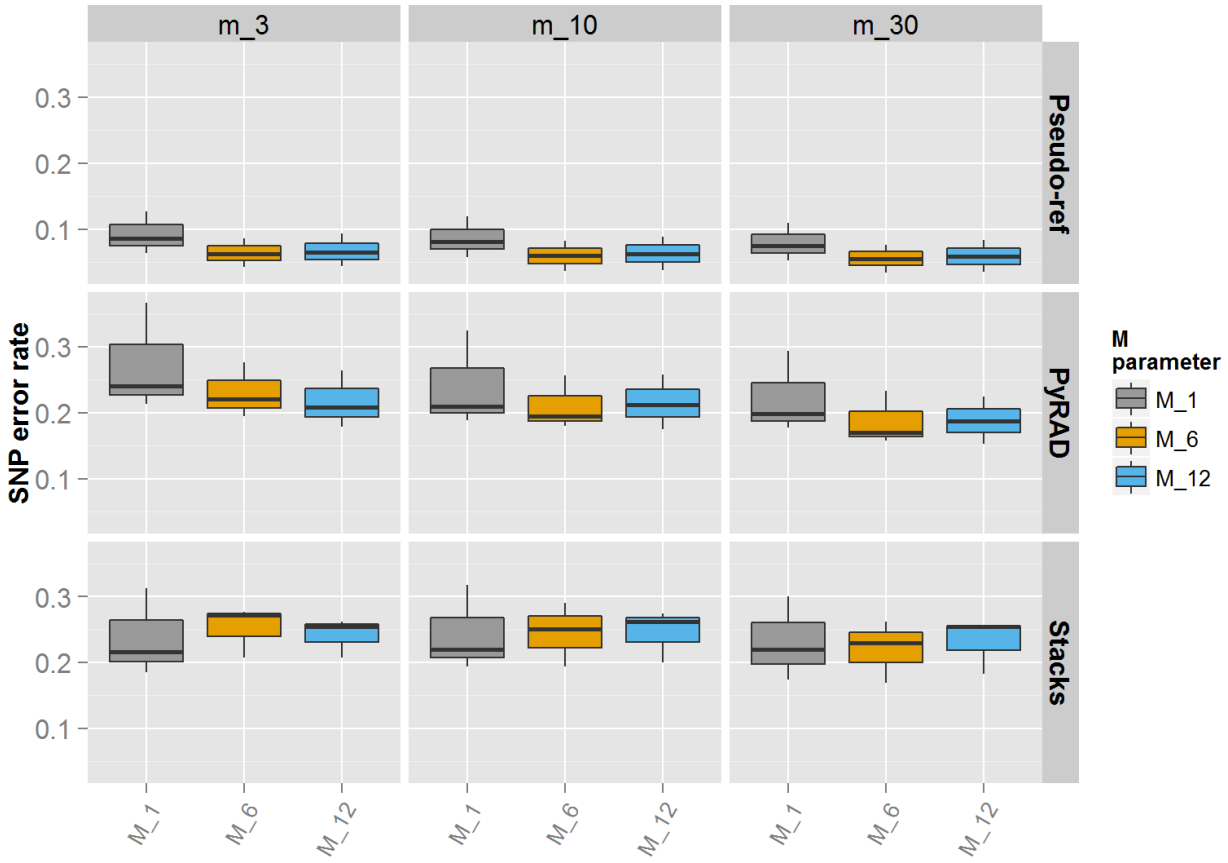


Figure 13. Comparison of the SNP error rates obtained from three different pipelines for a 'bad' quality RG at different combinations of m and M parameters. In the x-axes the different M parameter values tested. In the y-axes the proportion of markers showing a SNP error. Parameter m is the minimum coverage to call an allele and parameter M is the maximum number of nucleotide differences allowed between alleles.

From an optimization standpoint, and following the criteria we established in the Material and Methods section, for each pipeline we choose the combination of parameters that minimizes both SNP and Locus error rates, regardless of the number of loci or SNPs discovered. We did not follow the criteria suggested by Mastretta-Yanes et al. (2015) of searching for *de novo* assembly parameters that maximize the number of loci identified because in our dataset this adds redundancy (i.e., through the assembly of highly polymorphic RAD loci) and increases the risk of spurious assemblies due to paralogue over-merging. A trade-off between SNP and locus allele rates for bivalve RAD-Seq data was obtained at locus read-depth of 10 (m=10) and an allele mismatch of 6 (M=6). A summary of different statistics obtained for each pipeline is shown in Table 3. As can be observed, the pseudo-reference pipeline detects fewer RAD-loci than the PyRAD or Stacks pipelines, suggesting that low SNP error rates are caused by the elimination of spurious loci that potentially contain paralogues.

Table 3. Information content of the locus catalogue built with the optimized parameter values m=10 and M=6 using three *de novo* assembly pipelines.

		Stacks	PyRAD	Pseudo-ref
Chilean mussel	Number of RAD-loci	1,654	1,309	1,057
	Total number of SNPs	3,994	2,480	1,901
	SNP error rate	0.295	0.236	0.070
	Locus error rate	0.158	0.063	0.018
Blue mussel	Number of RAD-loci	4,117	3,485	2,328
	Total number of SNPs	12,616	9,794	5,562
	SNP error rate	0.030	0.033	0.006
	Locus error rate	0.091	0.073	0.034
Pacific Oyster	Number of RAD-loci	3,059	2,714	2,236
	Total number of SNPs	8,222	7,309	4,961
	SNP error rate	0.067	0.062	0.019
	Locus error rate	0.169	0.107	0.081

3.4 Conclusion

A critical step for RAD-Seq analyses of non-model organisms is *de novo* assembly. Even though *de novo* parameter settings have been shown to produce a significant variance in retrieved data (Catchen et al., 2013), technical replicates offer an opportunity for empirical optimization and are often preferred over other methods.

Here we used a set of technical replicates from three bivalve species – the British mussel, the Chilean mussel and the Pacific oyster – to optimize *de novo* assembly and genotype calling for this group of marine mollusks. Three different pipelines – Stacks, PyRAD and a pseudo-reference approach – were compared at different parameter settings. Although the strict comparison between pipelines is not feasible, at near analogous parameter settings, similar trends were observed. The assembly parameter that had the most significant impact in the amount of loci and SNPs retrieved was the one that controls the level of allelic divergence (parameter M in the text). The pipeline that showed both the lowest SNP and Locus error rates was the pseudo-reference approach. Two different variant calling programs were used for SNP calling (SAMtools and GATK-HC). Because most of the SNPs detected by GATK-HC were also detected by SAMtools and not vice versa, in future Chapters of this thesis only GTAK-HC is used for SNP calling. In our dataset, the best *de novo* assembly parameter profile (m=10 and M=6) gave a SNP error rate of 0.6% in the Blue mussel, 1.9% in the Pacific oyster, and 7% in the Chilean mussel technical replicate groups. These parameters values will be used to assemble the RAD-Seq data of Chapter 4.

A fundamental issue for genetic research in bivalves consists of being able to define unique genomic regions to extract meaningful genetic variation for analysis. However, in our exploratory analysis, a bimodal distribution of the coverage per locus histogram was observed (Figure 6). The unusual distribution was maintained under different combinations of parameter values test, suggesting it reflects a genomic property of bivalves (which is additionally supported by the normal unimodal distribution observed for the Nile tilapia control). The vast majority of polymorphic loci are observed at the peak with the higher average coverage per locus, suggesting that the abnormal bimodal distribution may be caused by (i) the extremely high levels of heterozygosity that prevents the correct assembly of particular loci or (ii) null alleles segregating at a high frequency in the data.

3.5 References

ANDREWS, K. R., HOHENLOHE, P. A., MILLER, M. R., HAND, B. K., SEEB, J. E. & LUIKART, G. 2014. Trade-offs and utility of alternative RADseq methods: Reply to Puritz et al. *Molecular Ecology,* 23**,** 5943-5946.

BAIRD, N. A., ETTER, P. D., ATWOOD, T. S., CURREY, M. C., SHIVER, A. L. & LEWIS, Z. A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One,* 3, e3376.

CATCHEN, J., HOHENLOHE, P. A., BASSHAM, S., AMORES, A. & CRESKO, W. A. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology,* 22**,** 3124-3140.

CATCHEN, J. M., AMORES, A., HOHENLOHE, P., CRESKO, W. & POSTLETHWAIT, J. H. 2011. Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes|Genomes|Genetics,* 1**,** 171-182.

CATCHEN, J. M., HOHENLOHE, P. A., BERNATCHEZ, L., FUNK, W. C., ANDREWS, K. R. & ALLENDORF, F. W. 2017. Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular Ecology Resources,* 17**,** 362-365.

CHUTIMANITSAKUN, Y., NIPPER, R. W., CUESTA-MARCOS, A., CISTUÉ, L., COREY, A., FILICHKINA, T., JOHNSON, E. A. & HAYES, P. M. 2011. Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. *BMC Genomics,* 12**,** 4.

CRUAUD, A., GAUTIER, M., GALAN, M., FOUCAUD, J., SAUNÉ, L., GENSON, G., DUBOIS, E., NIDELET, S., DEUVE, T. & RASPLUS, J.-Y. 2014. Empirical Assessment of RAD Sequencing for Interspecific Phylogeny. *Molecular Biology and Evolution,* 31**,** 1272-1274.

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T., SHERRY, S. T., MCVEAN, G. & DURBIN, R. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-2158.

DAVEY, J. W., CEZARD, T., FUENTES-UTRILLA, P., ELAND, C., GHARBI, K. & BLAXTER, M. L. 2013. Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology,* 22**,** 3151-3164.

EATON, D. A. R. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics,* 30**,** 1844-1849.

ETTER, P. D., BASSHAM, S., HOHENLOHE, P. A., JOHNSON, E. A. & CRESKO, W. A. 2011a. SNP Discovery and Genotyping for Evolutionary Genetics Using RAD Sequencing. *Methods in molecular biology (Clifton, N.J.),* 772**,** 157-178.

ETTER, P. D., PRESTON, J. L., BASSHAM, S., CRESKO, W. A. & JOHNSON, E. A. 2011b. Local De Novo assembly of RAD paired-end contigs using short sequencing reads. *Plos One,* 6, e18561.

FU, L., NIU, B., ZHU, Z., WU, S. & LI, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics,* 28**,** 3150-3152.

GONEN, S., LOWE, N. R., CEZARD, T., GHARBI, K., BISHOP, S. C. & HOUSTON, R. D. 2014. Linkage maps of the Atlantic salmon (Salmo salar) genome derived from RAD sequencing. *BMC Genomics,* 15**,** 166.

GRAHAM, C. F., GLENN, T. C., MCARTHUR, A. G., BOREHAM, D. R., KIERAN, T., LANCE, S., MANZON, R. G., MARTINO, J. A., PIERSON, T., ROGERS, S. M., WILSON, J. Y. & SOMERS, C. M. 2015. Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Molecular Ecology Resources,* 15**,** 1304-1315.

HARRANG, E., LAPEGUE, S., MORGA, B. & BIERNE, N. 2013. A High Load of Non-neutral Amino-Acid Polymorphisms Explains High Protein Diversity Despite Moderate Effective Population Size in a Marine Bivalve With Sweepstakes Reproduction. *G3: Genes|Genomes|Genetics,* 3**,** 333-341.

HEINRICH, V., STANGE, J., DICKHAUS, T., IMKELLER, P., KRÜGER, U., BAUER, S., MUNDLOS, S., ROBINSON, P. N., HECHT, J. & KRAWITZ, P. M. 2012. The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Research,* 40**,** 2426-2431.

ILUT, D. C., NYDAM, M. L. & HARE, M. P. 2014. Defining Loci in Restriction-Based Reduced Representation Genomic Data from Nonmodel Species: Sources of Bias and Diagnostics for Optimal Clustering. *BioMed Research International,* 2014**,** 9.

LI, H. 2009. The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics,* 25.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* 25**,** 1754-1760.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25, 2078-2079.

LI, W. & GODZIK, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 22**,** 1658-1659.

LOWRY, D. B., HOBAN, S., KELLEY, J. L., LOTTERHOS, K. E., REED, L. K., ANTOLIN, M. F. & STORFER, A. 2017. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources,* 17**,** 142-152.

MASTRETTA-YANES, A., ARRIGO, N., ALVAREZ, N., JORGENSEN, T. H., PIÑERO, D. & EMERSON, B. C. 2015. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources,* 15**,** 28-41.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research,* 20**,** 1297-1303.

MCKINNEY, G. J., WAPLES, R. K., SEEB, L. W. & SEEB, J. E. 2016. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources***,** 17, 656-669.

PANTE, E., ABDELKRIM, J., VIRICEL, A., GEY, D., FRANCE, S. C., BOISSELIER, M. C. & SAMADI, S. 2015. Use of RAD sequencing for delimiting species. *Heredity,* 114**,** 450-459.

PARIS, J. R., STEVENS, J. R. & CATCHEN, J. M. 2017. Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution,* 8**,** 1360-1373.

PETERSON, B. K., WEBER, J. N., KAY, E. H., FISHER, H. S. & HOEKSTRA, H. E. 2012. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLOS ONE,* 7**,** e37135.

PRYSZCZ, L. P. & GABALDÓN, T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research,* 44**,** e113.

RICHARDS, P. M., LIU, M. M., LOWE, N., DAVEY, J. W., BLAXTER, M. L. & DAVISON, A. 2013. RAD-Seq derived markers flank the shell colour and banding loci of the Cepaea nemoralis supergene. *Molecular Ecology,* 22**,** 3077-3089.

RICO, C., CUESTA, J. A., DRAKE, P., MACPHERSON, E., BERNATCHEZ, L. & MARIE, A. D. 2017. Null alleles are ubiquitous at microsatellite loci in the Wedge Clam (Donax trunculus). *PeerJ,* 5**,** e3188.

SAAVEDRA, C. & BACHÈRE, E. 2006. Bivalve genomics. *Aquaculture,* 256**,** 1-14.

SAUVAGE, C., BIERNE, N., LAPÈGUE, S. & BOUDRY, P. 2007. Single Nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster Crassostrea gigas. *Gene,* 406**,** 13-22.

SCHIRMER, M., D'AMORE, R., IJAZ, U. Z., HALL, N. & QUINCE, C. 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics,* 17**,** 125.

SHAFER, A. B. A., PEART, C. R., TUSSO, S., MAAYAN, I., BRELSFORD, A., WHEAT, C. W. & WOLF, J. B. W. 2016. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution***,** 8, 907-917.

SOVIC, M. G., FRIES, A. C. & GIBBS, H. L. 2015. AftrRAD: a pipeline for accurate and efficient de novo assembly of RADseq data. *Molecular Ecology Resources,* 15**,** 1163-1171.

TAKAHASHI, T., NAGATA, N. & SOTA, T. 2014. Application of RAD-based phylogenetics to complex relationships among variously related taxa in a species flock. *Molecular Phylogenetics and Evolution,* 80**,** 137-144.

TAKEUCHI, T., KAWASHIMA, T., KOYANAGI, R., GYOJA, F., TANAKA, M. & IKUTA, T. 2012. Draft genome of the pearl oyster Pinctada fucata: a platform for understanding bivalve biology. *DNA Research,* 19, 117-130.

TARIEL, J., LONGO, G. C. & BERNARDI, G. 2016. Tempo and mode of speciation in Holacanthus angelfishes based on RADseq markers. *Molecular Phylogenetics and Evolution,* 98**,** 84-88.

THIRIOT-QUIEVREUX, C. 2002. Review of the literature on bivalve cytogenetics in the last ten years. *Cahiers de Biologie Marine,* 43**,** 17-26.

THIRIOT-QUIÉVREUX, C., POGSON, G. H. & ZOUROS, E. 1992. Genetics of growth rate variation in bivalves: aneuploidy and heterozygosity effects in a Crassostrea gigas family. *Genome,* 35**,** 39-45.

TOONEN, R. J., PURITZ, J. B., FORSMAN, Z. H., WHITNEY, J. L., FERNANDEZ-SILVA, I., ANDREWS, K. R. & BIRD, C. E. 2013. ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ,* 1**,** e203.

WANG, S., MEYER, E., MCKAY, J. K. & MATZ, M. V. 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods,* 9**,** 808-810.

WILLIS, S. C., HOLLENBECK, C. M., PURITZ, J. B., GOLD, J. R. & PORTNOY, D. S. 2017. Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. *Molecular Ecology Resources,* 17**,** 955-965.

ZHANG, G., FANG, X., GUO, X., LI, L., LUO, R., XU, F., YANG, P., ZHANG, L., WANG, X., QI, H., XIONG, Z., QUE, H., XIE, Y., HOLLAND, P. W. H., PAPS, J., ZHU, Y., WU, F., CHEN, Y., WANG, J., PENG, C., MENG, J., YANG, L., LIU, J., WEN, B., ZHANG, N., HUANG, Z., ZHU, Q., FENG, Y., MOUNT, A., HEDGECOCK, D., XU, Z., LIU, Y., DOMAZET-LOSO, T., DU, Y., SUN, X., ZHANG, S., LIU, B., CHENG, P., JIANG, X., LI, J., FAN, D., WANG, W., FU, W., WANG, T., WANG, B., ZHANG, J., PENG, Z., LI, Y., LI, N., WANG, J., CHEN, M., HE, Y., TAN, F., SONG, X., ZHENG, Q., HUANG, R., YANG, H., DU, X., CHEN, L., YANG, M., GAFFNEY, P. M., WANG, S., LUO, L., SHE, Z., MING, Y., HUANG, W., ZHANG, S., HUANG, B., ZHANG, Y., QU, T., NI, P., MIAO, G., WANG, J., WANG, Q., STEINBERG, C. E. W., WANG, H., LI, N., QIAN, L., ZHANG, G., LI, Y., YANG, H., LIU, X., WANG, J., YIN,

Y. & WANG, J. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature,* 490**,** 49-54.

CHAPTER 4

Characterization of genome-wide patterns of segregation distortion in Pacific oyster full-sibling families

4.1 Introduction

Segregation distortion (SD) is the term used to describe a locus that shows a departure from the expected Mendelian genotypic ratios. The phenomenon of SD has been observed in a variety of species. However, it appears to be most prominent in plants, fungi and marine bivalves (Taylor and Ingvarsson, 2003, Plough, 2016, Saupe, 2012). Several endogenous and exogenous factors have been shown to cause the SD. Among the endogenous causes that may lead to SD, we find genetic elements that favor their transmission either by a meiotic drive (Lyttle, 1993) or by suppressing other gametes carrying a particular genotype (Engelstädter and Hurst, 2009). On the other hand, the main exogenous factor explaining SD in natural populations is natural selection. Natural selection can shift genotypic proportions at any stage of life, arising during the gametic phase or post-zygotically, during embryonic development or at a later stage in life (Falconer, 1975). The evolutionary importance of SD has been associated with the maintenance of species barriers in *Drosophila* (Tao et al., 2001) and *Arabidopsis* (Seymour et al., 2017). Additionally, SD has been shown to be a significant driving force behind sex-determination systems (Kulmuni et al., 2010, Taylor and Ingvarsson, 2003, Rutkowska and Badyaev, 2008).

The understanding of SD patterns in marine bivalves has lagged behind that of terrestrial species. The reason for this may be the paucity of efforts towards the development of genetic and genomic resources (Kelley et al., 2016), which is in line with their recent history of domestication (Duarte et al., 2007). Empirical evidence in pair-crosses of the Pacific oyster *Crassostrea gigas* and the European flat oyster *Ostrea edulis* suggested that SD originates early in life during larval development (Bierne et al., 1998, Launey and Hedgecock, 2001, Plough and Hedgecock, 2011). These studies originally aimed to unravel the basis of heterosis in bivalve mollusks and were therefore performed on inbred families. In this context, the strong selection against identical-by-descent homozygotes (homozygote disadvantage) was associated with a dominance (or associative overdominance) mechanism for heterosis. To explain the magnitude of SD observed by Launey and Hedgecock (2001), the wild founders of the inbred lines would have been required to carry 8-14 lethal equivalents in their genome.

If genetic inviability in oysters was caused by recessive deleterious alleles that became unmasked in an inbred genome, then genotype-dependent mortalities should be significantly reduced in randomly bred offspring. Following this premise, a subsequent

study aimed to map viability loci in an outbred crossing experiment (Plough et al., 2016). Contrary to expectation, they found that extensive SDs were mostly caused by selection against heterozygote genotypes in the offspring. In this new context, deleterious mutations were suggested to be partially dominant, leading to an overall reduction of the heterozygous genotype class in the progeny. Also, deleterious mutations were found to be segregating from both parents, with no predominant contribution of one parent over the other. The results of these experiments suggest that the interaction of spontaneous mutations and natural selection at early larval stages has a major role in shaping the SD patterns of marine mollusks (Plough, 2016). Despite SD being ubiquitous in bivalve species – with the proportion of distorted markers in a study ranging from 15-75% (see references in (Plough, 2016)) – the mechanisms underlying this phenomenon are incompletely understood.

To advance our understanding of SDs in marine bivalves, a more complete characterization of the genomic regions that show SDs is required. A SD region (SDR) can be identified by detecting clusters of loci within chromosome regions showing distortions. This because a locus responsible for SD is predicted to distort all linked markers (Luo and Xu, 2003, Zhou et al., 2015). With the advent of modern genomic technologies such as genotyping by sequencing and SNP arrays, a major opportunity exists to re-examine SD and its potential causes in bivalve species at a genome-wide scale, using thousands to tens of thousands of loci. An increase in marker density will allow detection of SDR in the bivalve genome at high precision, offering a valuable insight into SDs and their evolutionary significance.

In this Chapter, we examine the phenomenon of SD in bivalve shellfish by analyzing patterns of marker segregation in three Pacific oyster *Crassostrea gigas* full-sibling families. We applied two strategies for genome-wide SNP genotyping: a SNP-array (Gutierrez et al., 2017) and the RAD-Seq method (Baird et al., 2008, Miller et al., 2007). The purpose of using these two separate genotyping technologies was two-fold. First, RAD-Seq is a more accessible genotyping technique than a SNP-array, as no previous knowledge of the genome under study is required. Therefore, genomic resources for bivalve species are likely to expand based on the application of RAD-Seq or other reduced representation sequencing techniques (Robledo et al., 2017). However, bivalve genomes show a remarkably high level of sequence polymorphism and genome complexity (Guo et al., 2015, Zhang et al., 2012), which may influence the genotyping outcome depending on the

technology applied (Lemer et al., 2011). In bivalves, we anticipate the RAD-Seq approach to be more prone to genotyping artifacts compared to an array, although the magnitude of this difference is uncertain. By comparing the frequency of Mendelian inconsistencies in nuclear families genotyped with both platforms, we assess the suitability of RAD-Seq for performing genetic analyses in a species with a complex genome.

And second, we map the SD loci from both technologies (SNP-array and RAD-Seq) to the available Pacific oyster reference genome assembly (Zhang et al., 2012). By co-locating the data, we identified high confidence SDRs, as they showed evidence of distortions in markers from both genotyping platforms. A subset of scaffolds was then selected to represent the ten Pacific oyster linkage groups (2n=20). A high resolution map of the SDRs was obtained with the SNP chip data, which allowed a more accurate inter-family comparison in contrast to previous similar attempts that used a lower marker density (Jones et al., 2013, Hedgecock et al., 2015). Also, we examine the SDRs to determine, when possible, whether SD in the Pacific oyster progeny is associated with the paternal or maternal alleles. The results presented here provide insight into the frequency, pattern, and nature of recurrent segregation distortion in a marine bivalve species of ecological and economic relevance, the Pacific oyster.

4.2 Material and Methods

4.2.1 Oyster families

In 2016, one full-sib (family 19) and two half-sib families (families 29 and 30) derived from pair-crosses of Pacific oysters (*C. gigas*) were created and reared at the Centre for Environment, Fisheries and Aquaculture Science (Cefas), Weymouth, the UK. These families were part of an experiment designed to identify regions of the oyster genome responsible for susceptibility/resistance to a variant of the oyster herpes virus (OsHV-1 µvar) (see Material and Methods Chapter- section 2.1.1.3). Families were tracked throughout the experiment to ensure no mixing of individuals. Tissue samples from the parents were preserved in absolute ethanol. Offspring samples (~2-8 months old) consisted of whole individuals that were snap frozen either when they died during the experiment, or as survivors at the end of the experiment. In total, three nuclear families – two of which shared the same dam (families 29 and 30) – consisting of the parents and between 40-47

offspring each, were sampled and transferred to the Roslin Institute, Edinburgh, the UK, for DNA extraction, genotyping and analysis.

4.2.2 DNA extraction and quantification

DNA extractions and library preparation methods were followed according to the protocol outlined in section 2.1.3 of Material and Methods (Chapter 2). DNA samples were quantified with a Qubit® 2.0 instrument (Invitrogen) and were diluted to a working concentration of 25ng/ul.

4.2.3 Molecular marker genotyping

4.2.3.1 SNP chip

All oyster samples - three families comprising 143 individuals in total – were genotyped with a 55 K Affymetrix SNP array developed by our group (Gutierrez et al., 2017). The SNP chip was developed for two oyster species, the Pacific oyster (*C. gigas*) and the European flat oyster (*Ostrea edulis*). Consequently, only the Pacific oyster probes of the array (40 K SNPs) were scanned for genotypes. The genotyping of the samples is described in detail in Gutierrez et al. 2017.

4.2.3.2 RAD sequencing

The genomic DNA from a subset of the same three Pacific oyster families genotyped with the array (n total = 78) were used for genome-wide SNP discovery and genotyping using RAD-Seq. Six sub-libraries were prepared following the *SbfI* RAD-Seq protocol described in the Materials and Methods Chapter. The quality of each sub-library was assessed on an Agilent 2100 Bioanalyzer. Libraries were pooled in equimolar amounts and adjusted to a final concentration of 10nm. The library was sequenced on a single lane of an Illumina HiSeq2500 instrument (125bp paired end reads) by the Edinburgh genomics facility, University of Edinburgh, the UK.

4.2.4 Pre-processing of RAD-Seq data

Paired end reads were de-multiplexed using the unique barcode combination assigned to each sample, and reads with low-quality scores (QC<10; Catchen et al. (2011)) or ambiguous barcodes were discarded. Due to the fact that we used a single-end *de novo* assembly, PCR duplicates were removed, earlier in the analysis from the paired-end fastq files using the BBtools dedupe.sh module. A preliminary analysis showed an enrichment of SNPs towards the 3' end of the single-end reads (Figure 2, Chapter3). Therefore, 20bp were trimmed from both reads in a pair with the fastx_toolkit v 0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/). The number of reads per individual ranged from 51,264 to 1,634,309, with an average of 623,460 and standard deviation of 333,971. After the removal of individuals with low number of quality reads, and due to the significant variation in read count across samples, we normalized samples to the read number of the sample with fewer read counts (~320,000 reads). This approach was taken to facilitate the direct comparison across samples when coverage is taken into consideration to explore the presence of null alleles in the RAD-Seq data (Chapter 5).

4.2.5 De novo assembly of RAD loci

De novo assembly was performed using the pseudo-reference pipeline developed and evaluated in the previous Chapter (section 3.2.5, Chapter 3). Briefly, *de novo* assembly of Pacific oyster single-end reads was performed using Stacks (Catchen et al., 2011, Catchen et al., 2013) with the following assembly parameters: reads were assembled into RAD-loci with the *ustacks* module using a minimum stack depth of 10, and a maximum number of mismatches between stacks (or the two putative alleles of an individual) of 6. In addition, we allowed for the presence of indel variation enabling a '--gapped-alignment'. The consensus sequence of the RAD loci defined for each individual were extracted, grouped by family, and collapsed into a representative set of family-specific loci by applying an 94% sequence identity threshold in the software CD-hit (-c 0.94) (Li and Godzik, 2006, Fu et al., 2012). These files were used to generate a pseudo-reference genome for each family, to avoid reference bias and increase the sensitivity (see Chapter 3). Single-end reads from all samples were aligned to their respective family pseudo-reference using BWA (Li and Durbin, 2009) utilizing parameters that allow for a high divergence between the sequenced

samples and the reference (full BWA command: bwa aln -l 210 -o 2 -n 0.01 -e 10 -d 12). Ambiguously mapped reads were removed from the dataset.

4.2.6 Variant calling

SNP calling on the mapped reads were performed utilizing GATK-HC (HaplotypeCaller) at default parameters, as described in Chapter 3 section 3.2.5. Variant discovery was conducted in each of the three oyster families separately, given that each family was mapped to an independent set of family-specific pseudo-reference loci. Subsequently, hard-filters were applied, following recommendations for variant calling in non-model species. Variants were filtered for a QualityByDepth (QD) >5. QD represents the normalization of QUAL score (confidence) of a variant by allele depth, and it was applied to avoid the inflation of the variant quality because of a deeper coverage. Loci required a minimum read depth of 10 for variants within the loci to be retained for further analyses. Most of the reads had a mapping quality of 37 due to previous filtering, therefore, no mapping quality filter was applied at this stage of the analysis. No strand bias filter was applied because of the nature of the RAD-Seq method. Markers with missing data for more than 10 % of individuals in a family were excluded with vcftools (Danecek et al., 2011). The visual inspection of RAD-loci (= 100 bp length) harboring extreme SNP polymorphisms showed evidence of the collapse of paralogous loci. This result was anticipated from the flexible *de novo* assembly parameters used in this study (M=6) and the highly variable genome of *C.gigas.* Nevertheless, we purposely attempted to collapse highly similar paralogous loci to facilitate their filtering from the dataset, as they will typically exhibit more than the expected two haplotypes for a diploid species. Conflicting, paralogue-containing RAD-loci were removed from the dataset with a script that filters based on haplotype number (Willis et al., 2017). Initially, and to avoid sampling physically linked markers, we selected RAD-loci that contained a single SNP marker. However, loci that meet this requirement represent less than half of the data (46%). Therefore, using this approach would considerably reduce the resolution of the segregation analysis. Instead, to maximize the amount of genetic information, we constructed SNP haplotypes from the RAD-Seq data (Willis et al., 2017) and performed a haplotype segregation analysis.

4.2.7 Segregation analysis

4.2.7.1 Individual marker segregation analysis

As a first examination of SD, we performed a single-marker segregation analysis with the genotypes obtained for the three oyster families using the (i) SNP markers derived from the SNP array, and (ii) haplotype markers, which were coded from the RAD-Seq data. Family genotype data obtained from both genotyping platforms were classified into five segregation types according to parental genotypes (sire x dam or vice versa: "AA x AA", "AA x BB", "AB x AA", "AB x AB", "AB x BB"). Significant deviations from Mendelian ratios were tested using a chi-square test, with the significance levels (α=0.05) adjusted using a Benjamini-Hochberg (B-H) procedure (Benjamini and Hochberg, 1995). When both parents are heterozygous for the same alleles ("AB x AB") the progeny genotypes were tested against a 1:2:1 (homozygous: heterozygous: homozygous) ratio. When only one parent was heterozygous ("AB x AA"), the expectation is 1:1 (homozygous: heterozygous) for genotype frequencies.

4.2.7.2 Genome-wide patterns of segregation

To characterize the genome-wide patterns of SD in the Pacific oyster, the available reference genome (GCA_000297895.1), which is not yet assembled into chromosomes, was used to colocate the RAD-loci and the SNP array probes onto genome scaffolds. A genome scaffold is an incomplete reconstruction of a portion of the genome composed by contigs (i.e., contiguous length of DNA sequence in which the order of bases is known with high confidence) and gaps (Barton and Barton, 2012). The available Pacific oyster reference genome (Zhang et al., 2012) comprises approximately 7,500 scaffolds (N50 = 401,585) and approximately 30,000 contigs (N50 = 31,239) that cover 87% (557 Mb) of the oyster genome size estimated from flow cytometry (Zhang et al., 2012). A nucleotide database containing all scaffolds + contigs from the Pacific oyster assembly was created using BLAST (Altschul et al., 1990). The RAD-loci (100bp) and Affymetrix probes (70 nucleotides in length) were aligned against the database using 100 % of query coverage and an E-value cut-off of $1E^{-50}$. The expectation value (E-value) is a parameter in BLAST that describes the number of hits expected to be seen by chance when aligning against a database of a particular size (https://blast.ncbi.nlm.nih.gov/). Consequently, at lower E-values, hits are

more significant. To avoid assemblies that contained paralogues or repetitive elements, only queries that aligned uniquely to the database were retained for further analysis. The scaffolds that had segregation information for both genotyping technologies were ranked by the total number of markers showing SD, and the top ten were selected for plotting the –log10 P-values for each segregation analysis. Nevertheless, by cross-referencing these scaffolds with a second generation Pacific oyster linkage map (Hedgecock et al., 2015), we noted that most of the scaffolds with highly significant SD (5 out of 10) mapped to a single linkage group (LG), LG 10. Since we are interested in obtaining a general genomic view of SDs instead of focusing on a single LG, we kept the five scaffolds that mapped to LG 10 for a high-resolution analysis and selected additional individual scaffolds (one from each LG) to examine patterns of distortion on sections of the remaining chromosomes. For identifying and delimiting SDRs, only SNPs from the array were used, due to their higher density. Because biological segregation distortion (i.e., not caused by genotyping errors) is always associated with clusters of distorted markers, we declared a SDR when, at least in a single family, a minimum of two consecutive distorted SNP markers was observed. Once SDRs were detected with the array markers, we utilized the RAD-Seq haplotypes to evaluate whether the distortions in the offspring originated from unexpected ratios associated with specific paternal or maternal alleles. This approach was used because haplotypes are expected to be more informative than (typically bi-allelic) SNP markers, allowing us to potentially distinguish parent-of-origin allelic effects. For each informative haplotype cross (e.g., "AB x CD") we test for the deviation from 1:1 segregation in the male and female gametes independently using a chi^2 goodness of fit test at a 5% significance level.

4.3 Results

With the SNP-array, a total of ~25,000 markers were successfully genotyped in each of the three oyster families. The total genotyping call rate was > 99 % across individuals. The number of segregating SNPs was similar across families, with an average of 9,766 informative markers per cross (SD=33)

Regarding the RAD-Seq approach, the sequencing of 78 Pacific oyster individuals yielded a total of 200 million reads. Reads were quality-filtered and separated by family to perform a family-based *de novo* assembly of RAD-loci as described in Chapter 3 (section 3.2.5). The

final pseudo-reference contained 4,183 RAD-loci for family 19, 3,324 for family 29, and 3,055 for family 30. RAD-loci that contained more than two haplotypes – around 420 from each family – were removed from the dataset, as they likely result from the overmerging of paralogues. Compared to the SNP-array, the number of SNPs genotyped with the RAD-Seq method was lower and varied considerably across families. After SNP calling and filtering, we identified 4,806 SNPs in family 19, 3,340 in family 29, and 2,975 in family 30. The average SNP frequency of the RAD-loci corresponded to 1 every 48bp, and was similar across families. After haplotyping nearby SNP markers, informative parental cross types (e.g., "AA x AB" or "AB x AC") from families 19, 29 and 30 comprised 1,041, 962 and 968 haplotype markers, respectively.

4.3.1 Segregation analysis

4.3.1.1 SNP chip

Over 54 % (41,373 markers) of the SNPs genotyped across families were fixed for the same allele in both parents, and were therefore discarded from the segregation analysis. The most frequent informative parental cross found across families was of the "AB x AA" type (or vice versa for homozygous sire, "AA x AB"), representing 78% of the markers yielding segregation data. This parental cross also contains the greatest number of distorted loci (Table 1). Across the three families, the total number of significantly distorted markers after multiple correction was 3,260 (33%), 1,148 (12%) and 2,229 (23%) for oyster families 19, 29 and 30, respectively. Generally, the pattern of segregation distortion reported in the literature for bivalves is a deficiency of heterozygote genotypes; although heterozygous excess has also been reported, particularly with microsatellite markers in partially inbred oyster families (Bierne et al., 1998, Launey and Hedgecock, 2001). However, for the distorted markers detected in this study, no clear bias for or against the heterozygous genotype state was observed. Only 22% of the distorted "AB x AB" crosses were caused by a deficiency of heterozygote genotypes in the offspring; the remaining distortions were caused by a bias against homozygous genotypes in the offspring. Specifically, when SD was caused by deficiencies of homozygotes (n total = 498 markers), 54% (271) of the instances were because of the complete absence of one of the expected homozygote categories, and 15% (73) were represented by cases in which both expected homozygote categories are not

observed in the offspring. For the "AB x AA" segregation type (where one of the parents is heterozygous and the other homozygous), approximately half of the SNPs (51%) that showed deviation from the expected Mendelian ratios were caused by a deficiency of heterozygous genotypes, with the other half caused by an excess of heterozygous genotypes. A more detailed evaluation of patterns of heterozygote deficiencies or excess by type of parental cross (sire x dam or vice versa: "AB x AA" or "AA x AB") is discussed below (see Results section: Patterns of marker segregation across Pacific oyster genome scaffolds).

4.3.1.2 RAD-Seq method

Initially, we carried out a segregation analysis of SNP markers selected from RAD-loci with low polymorphism (i.e., loci that harbored only one SNP). However, this approach significantly reduced the number of markers interrogated, as most of the RAD-loci (around 54%) harbored more than one SNP. We therefore opted to perform a haplotype segregation analysis. Among the five categories of parental mating types identified, the majority of informative parental crosses were heterozygous in one parent and homozygous in the other (i.e., "AB x AA") (Table 1). The segregation analysis of haplotype markers across families indicated that, among the informative crosses, 360 haplotype-markers (35%) were significantly distorted in family 19, 140 (15%) in family 29, and 220 (23%) in family 30. Similar to the array data, and as expected given the same families were genotyped, departures from Mendelian proportions did not show a clear tendency against a certain genotype category. Distortions were caused in 54% and 19% of the instances by a deficiency of a certain heterozygous haplotype combination in the offspring for "AA x AB" and "AB x AB" crosses respectively.

Overall, and given the difference in the number of markers interrogated by both platforms (~25,000 SNPs vs. ~990 haplotypes), results for the segregation analysis in three oyster families were generally consistent between genotyping methods. The proportion of distorted markers identified by the SNP array and the RAD-Seq method was stable across families. Also, distorted SNPs or haplotypes were consistent in that both were not skewed towards a specific direction (e.g., towards the typical deficiency of heterozygote genotypes observed for bivalve species). Nonetheless, one important distinction was the higher

number of apparent Mendelian errors identified in the RAD-Seq data, which resulted in a greater loss of data because of quality control pruning; 8% of SNP markers were removed from the SNP array data due to Mendelian errors and 27% of RAD loci were removed from the RAD-Seq data.

Table 1. Mendelian segregation results of markers genotyped in three Pacific oyster families using two high-throughput genotyping methods. Numbers in parenthesis in the body of the table represent the number of markers with significant deviations at the nominal α=0.05**.**

	SNP markers (SNP array)		**Haplotype markers (RAD-seq)**	
Mating type	**N^o**	**SD loci**	**N^o**	**SD loci**
AA x AA	41,373	-	266	-
AA x BB	3,624	-	674	-
AB x AA	24,274	6,292 (8,839)	2,644	389 (625)
AB x AB	6,737	635 (1,088)	1,136	331 (496)

4.3.2 Detection of Mendelian errors across genotyping platforms

The term Mendelian error is used in this study to describe occasions in which an offspring carries an allele that does not fit with the expectations of Mendelian inheritance based on the parental genotypes. The term is used irrespective of the source of the error. It may be caused by technical or biological factors such as genotyping errors (e.g., null alleles), erroneous family assignment, or, in very rare cases, *de novo* mutations.

4.3.2.1 SNP array

We observed an association between the type of parental cross and the number of Mendelian errors. The category that proportionally showed the highest level of Mendelian inconsistencies was the one for which SNPs were fixed for alternative alleles in the parents (i.e., "AA x BB"). Levels of inconsistencies in this category were similar across families and affected from 28 - 31% of the SNP markers (Figure 1). It should be noted that the probability of observing true Mendelian errors in this type of parental cross is higher than in any other cross, given only one genotype category is expected in the offspring (100% AB). Taking into consideration all the informative SNP matting types across families (n total = 34,635 segregation ratios), we estimated that 8% had Mendelian inconsistencies. This apparently high value is similar to the 7.4% reported by Qi et al. (2017) for a Pacific oyster family genotyped with a high-density 190K SNP array. Most of the SNPs showing Mendelian errors did not overlap across families (only 61 SNP markers were distorted across the three families), suggesting that errors are not associated with the array design (e.g., due to paralogous sequence variation). The pairwise comparison between families indicates that the majority of distorted SNP markers are exclusive to a family, instead of distorted markers being shared by two families (Table 2). It must be noted however that this value depends on the type of parental cross (i.e., genotypes of the parents), thus limiting a reliable cross-comparison across families.

Table 2. Number of shared and unique distorted SNP markers between families. Only significantly distorted markers after B-H correction are reported.

Comparison (Family A vs Family B)	Unique to Family A	Unique to Family B	Present in both families
Family 19 vs 29	3,368	993	155
Family 19 vs 30	3,027	1,760	496
Family 29 vs 30	838	1,946	310

4.3.2.2 RAD-Seq method

The number of Mendelian errors for the RAD-Seq haplotypes is markedly higher for all mating types, although we find a similar trend (Figure 2). The "AA x BB" category, which represents 15% of the observed parental crosses (whether informative or not), proportionally shows the highest number of Mendelian errors – 86%. The remaining informative category "AB x AA" reached 30% of Mendelian errors, although a strict comparison of this category to the equivalent category in the SNP array is not possible. This is because for the haplotype analysis the category comprises any cross for which one parent is heterozygous and the other homozygous, independent if they share a haplotype allele or not. For example, both "AB x CC" and "AB x BB" mating types are pooled into the nominal "AB x AA" category.

By genotyping the same three oyster families with two genotyping platforms, we were able to detect Mendelian errors and compare these estimates between platforms (SNP array vs. RAD-seq method). The SNPs genotyped with the array were more robust compared to those obtained from RAD-seq. This observation is supported by the proportionally lower number of Mendelian errors observed for the array-derived SNPs (Figure 1). By definition, a Mendelian error is caused by an unexpected genotype in the progeny. These unexpected genotypes in our dataset are not novel in the sense that the same alleles observed in the parents are present in the offspring but in a different combination than expected under normal inheritance. Typically in bivalve genetic research, these unusual offspring genotypes (which generate the Mendelian errors) are attributed to null alleles and are recoded when possible (Launey and Hedgecock, 2001, Plough et al., 2016). For example, the parental cross "AB x AA" that shows three genotype categories in the offspring (AA, AB, and BB) may be recoded as "AB x AB", based on the assumption that the true parental genotype is in fact "AB x AØ" (with Ø indicating a null allele). Instead of recoding or removing the markers that showed evidence of Mendelian errors, in the current study we aimed to study their potential origin in Chapter 5.

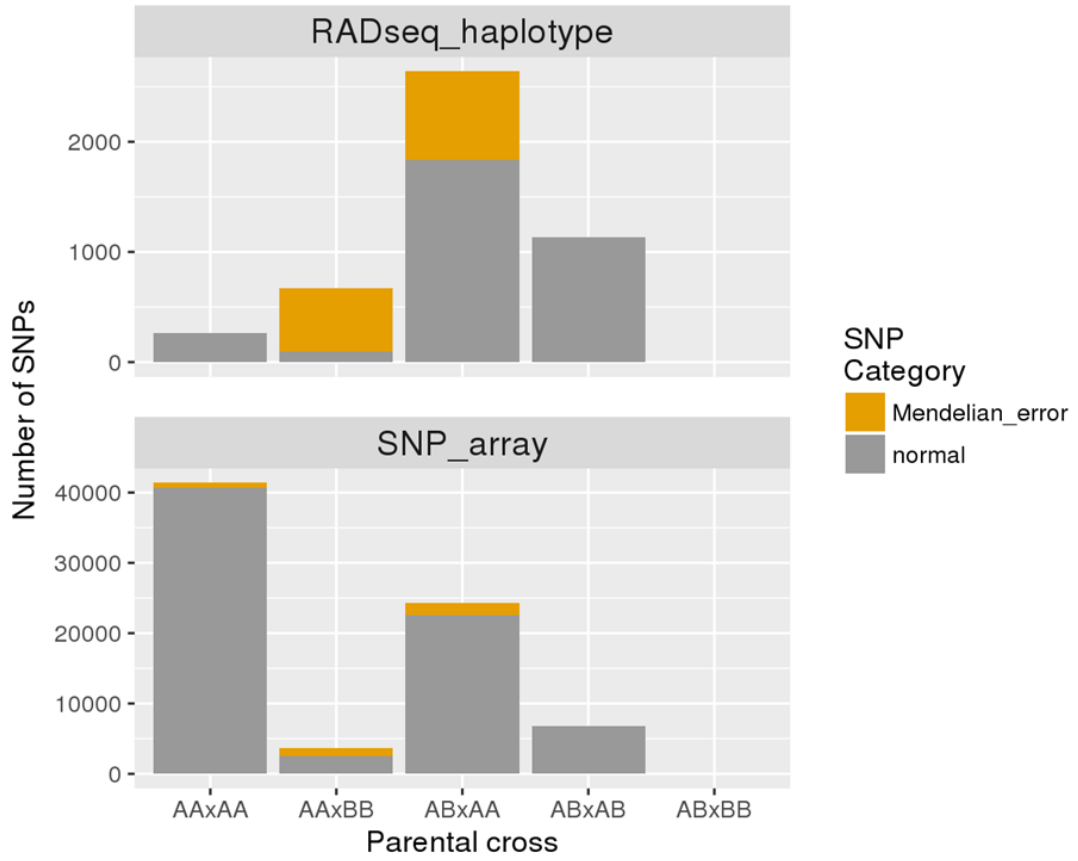


Figure 1. Comparison between the segregation analysis of the RAD-Seq-derived haplotype markers (top) and SNP markers genotyped with a SNP array (bottom). In the x-axes the different types of the parental cross are shown, while in the y-axes the number of markers per category are presented. For each mating type, the gray fraction of the bars represents the number of cases in which the marker segregates according to Mendelian rules, whereas the orange fraction represents instances in which the marker segregation data is inconsistent with the Mendelian law of inheritance (Mendelian error).

4.3.3 Patterns of marker segregation across Pacific oyster genome scaffolds

We compared the results from the segregation distortion analysis obtained with the SNP chip and the RAD-Seq method by anchoring the data to the available Pacific oyster reference genome. Genotype concordance between platforms was not tested at the individual marker level, as variants detected by each platform did not overlap. The flanking sequence of the SNPs interrogated in the array (i.e., probes) and the RAD-loci were aligned against a custom database containing all assembled scaffolds and contigs (7,658 and

30,459, respectively) using an e-value of $1E^{-50}$. Only queries with one match to the database were kept for analysis, in order to avoid paralogous regions. All probes (~40,000) from the SNP array aligned to the reference genome; as expected, considering that the *C. gigas* portion of the array was designed from SNPs detected after the alignment of WGS data to the reference genome (Gutierrez et al., 2017). Around 91 % of the probes aligned uniquely to either a contig or a genome scaffold. On the other hand, the RAD-loci that were assembled *de novo* within each family aligned to a total of 612 contigs and 2,483 scaffolds. Compared to the array, a lower mapping rate was found for the RAD-Seq loci, which ranged from 68% to 73%, depending on the family. Interestingly, we also find variability in the number of exact nucleotide positions interrogated by the RAD-Seq method. Theoretically, the same genomic regions should have been sampled across the individuals belonging to the three Pacific oyster families, which are defined in our study by the *SbfI* recognition site. Consequently, if we find a scaffold position to which only loci from one family are aligned, we would interpret it as a (genome) sampling error that may be of biological or technical origin. To test the sampling efficacy of the RAD-Seq method, we searched for the exact match of the family-specific RAD-loci to a scaffold position, independent of whether the locus is monomorphic or polymorphic. At this point of the study, contigs were removed from the reference assembly because of their shorter length, which may limit further comparative analysis. By placing the family-specific RAD loci onto scaffolds, we noted that, although most of the genome positions were sampled across families (52% → 2,915 out of 5,614 RAD-loci), a significant amount were exclusive to either a family or pair of families. RAD-loci identified exclusively in family 19 represented 20% of the data (1,109 out of 5,614 loci). Additionally, families 29 and 30 – which had the same dam – also shared a significant number of loci (13% → 729 from 5,614) that were not sampled, for unknown reasons, in the individuals from family 19. In contrast, the SNP-array platform achieved a higher technical consistency, as the same genomic coordinates, and therefore scaffolds were interrogated across families.

To evaluate genome-wide SD patterns, we selected 228 scaffolds that contained segregation information for all families on both genotyping platforms, allowing a joint analysis. Nearly 67 % (153/228) of these scaffolds contained SNP markers that showed congruent results between genotyping platforms. That is, on these scaffolds, markers from the RAD-Seq method and the SNP array both show evidence of distortion (93% of the cases), or vice versa, both agree on the absence of SD (7% of the cases). For the instances

where the segregation analysis results disagreed between genotyping platforms (33% → 75/228 scaffolds), it was caused mostly by the observation of distortion in the SNP chip genotyping data and absence of significant distortion in the RAD-Seq data. The latter result is probably consistent with the fact that, because of higher marker density, the SNP chip can detect more distortions than the RAD-Seq methodology.

As mentioned above, 153 scaffolds contained segregation information for all families and additionally showed an agreement of the SD results from the two high-throughput genotyping platforms. Because at this stage the aim was to obtain a fine resolution of SDRs across the Pacific oyster genome, only the genotyping data from the SNP array was used. In a previous study, a significant amount of errors in the scaffold assemblies of the Pacific oyster reference genome were detected (Hedgecock et al., 2015). We therefore removed potentially misassembled scaffolds from our analysis by choosing those that mapped unambiguously to a linkage group in the study performed by Hedgecock et al. (2015). The Pacific oyster has 10 LGs, however, unambiguous mapping of scaffolds was only possible for eight LGs (LGs two and four were excluded). To detect SDRs at a high resolution, the scaffold with the largest number of distorted loci among the total number of scaffolds mapped to a specific LG was selected for plotting the -log10 p-values obtained for the segregation analysis. Given we noted that a significant amount of distortions were observed on scaffolds that unambiguously mapped to LG 10, we additionally performed a high-resolution analysis on this LG. The length of the selected reference scaffolds ranged between ~0.3 and ~1,4Mb, with an average of 0.8 Mb.

The genome-wide patterns of SD over a theoretically representative set of regions of the Pacific oyster genome are presented in Figure 3. Non-segregating markers are omitted (e.g., "AA x AA" type crosses). The pattern of SDRs was not random, as 8 out of 12 scaffolds showed consistently distorted regions in families 19 and 30 (e.g., LG 7 in Figure 3). Among the three Pacific oyster families, family 29 showed fewer distorted regions, with only four scaffolds affected (scaffold number 315, 1,179, 1,866 and 1,528). Across families, we found that 90 % of the SNPs displaying SD across the scaffolds belonged to the parental crosses in which one parent was heterozygous and the other homozygous (sire x dam; "AA x AB" or "AB x AA"). A significant association between parental mating type and the number of distorted SNPs was found (chi^2 test of independence: p-value<0.001), indicating that it is more likely to observe a distortion if the parental cross is a het vs. hom (or vice versa). If we

only consider the two families that show the highest consistency regarding SDRs (i.e., families 19 and 30), we observe that, overall, for crosses were the sire is homozygous and the dam heterozygous ("AA x AB"), most of the distortions are caused by a deficiency of homozygous genotypes in the offspring (goodness-of-fit test; p-value<0.001), as opposed to a deficiency of heterozygote genotypes. Whereas for the cases in which the sire is heterozygous and the dam homozygous ("AB x AA"), SDs were equally as likely to be caused by a deficiency of the homozygote or heterozygote genotype categories (goodness-of-fit test: p-value=0.7). An illustration of the tendency of the distortions at SDRs shared between families 19 and 30 is shown in Figure 4.

Based on the segregation analysis of SNP markers covering 8 LGs, we find that SDRs are extensive in nature, as they usually affect a high proportion of the scaffold lengths, typically covering from 30% to 90% of the scaffold (Table 3). Since most distorted regions occurred towards the end of the scaffolds, we cannot define their limits. The consistency observed between the patterns of SDs of families 19 and 30 suggests that these regions are biologically meaningful, and may comprise viability related genes. A striking feature of SDs can be observed when the frequency of the minor allele of distorted markers (estimated from the progeny) is plotted along the commonly distorted regions in families 19 and 30 (Figure 4). A ragged pattern of allele frequencies is observed along the SDRs of both families, which shows more than 2-fold differences in variation between regions in close physical proximity. This pattern of nearby SNP markers showing mixed directions of distortion (excess and deficiency of heterozygote genotypes), along with symmetrical minor allele frequencies around the expectation of 0.25 (dotted line in Figure 4), suggests that markers are linked in large blocks.

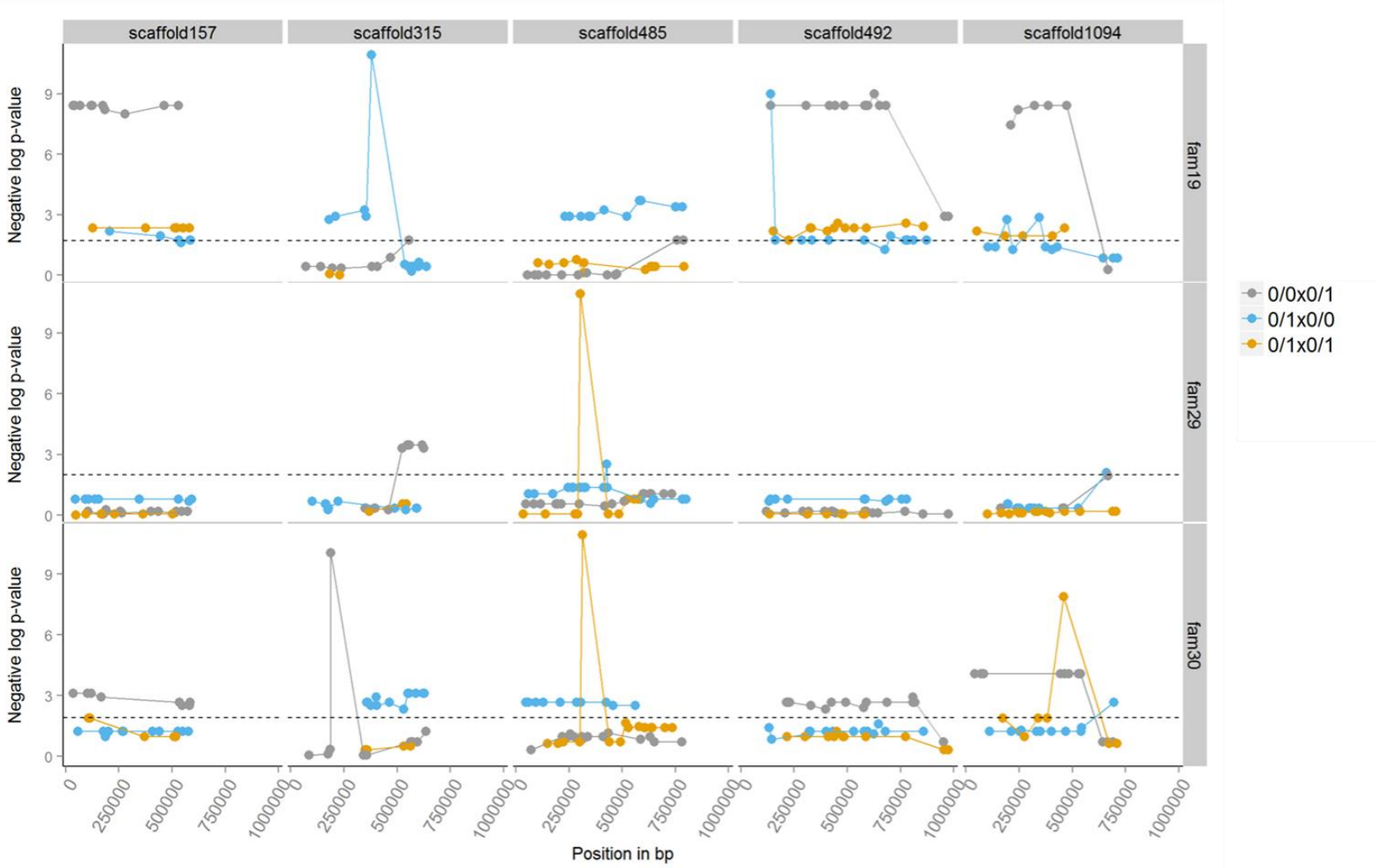
Additional information about the genes underlying these SDRs was obtained from the annotated Pacific oyster genes available in the ENSEMBL database (Yates et al., 2016). Most of the SDRs comprise a high number of genes, from 8 to 62, with an average of 27 genes. This broad set of candidate genes prevents a deeper understanding of the underlying mechanism by which SD may arise in bivalves.

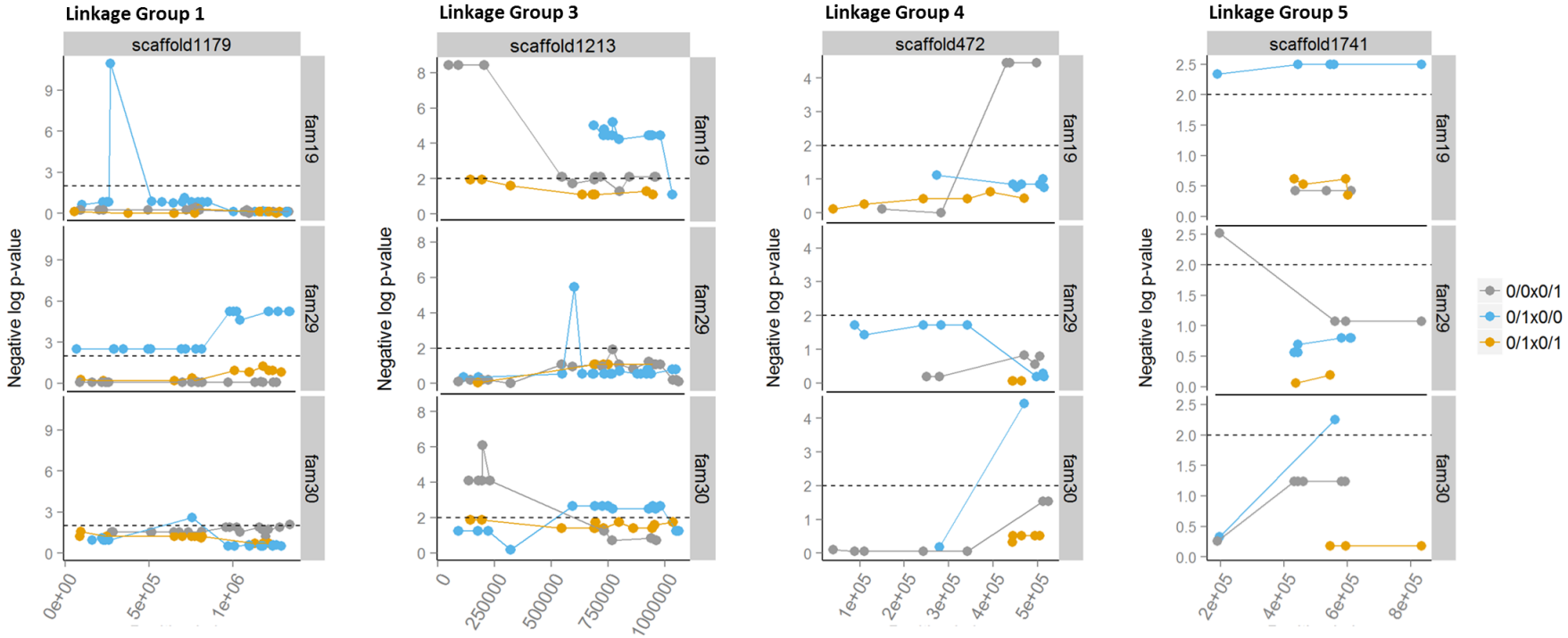
Table 3. SDRs detected in three Pacific oyster families after B-H correction (p<0.05).

Family ID	SDR (marker interval in bp)	SDR (length in bp)	LG	Number of SD markers	Distortion toward parent
Family 19	25,998-526,367	500,369	10 (sc. 157)	10	dam
	173,169-374,949	201,780	10 (sc. 315)	5	sire
	223,833-780,350	556,517	10 (sc. 485)	12	sire
	131,172-973,908	842,736	10 (sc. 492)	14	dam
	203,554-469,774	266,220	10 (sc. 1094)	5	dam
	47,516-981,710	934,194	3	20	sire & dam
	430,759-498,217	67,458	4	3	dam
	189,520-834,114	644,594	5	5	sire
	215,864-399,160	183,296	6	7	dam
	67,538-385,984	318,446	7	16	dam
	36,236-282,200	245,964	8	5	dam
Family 29	530,025-628,588	98,563	10 (sc. 315)	5	dam
	68,232-1,339,292	1,271,060	1	19	sire
	197,505-435,518	238,013	6	9	dam
	36,236-173,241	137,005	8	5	dam
Family 30	34,507-583,841	549,334	10 (sc. 157)	8	dam
	357,800-628,588	270,788	10 (sc. 315)	12	sire
	46,647-561,639	514,992	10 (sc. 485)	10	sire
	214,092-818,006	603,914	10 (sc. 492)	13	dam
	42,184-539,053	496,869	10 (sc. 1094)	8	dam
	137,156-981,719	844,563	3	14	sire & dam
	391,901-435,518	43,617	6	5	sire
	257,022-563,100	306,078	7	12	sire

sc: scaffold

Linkage Group 10




Linkage Group 1
scaffold1179
Linkage Group 3
scaffold1213
Linkage Group 4
scaffold472
Linkage Group 5
scaffold1741
fam19
fam29
fam30
Negative log p-value
0e+00
5e+05
1e+06
0
250000
500000
750000
1000000
1e+05
2e+05
3e+05
4e+05
5e+05
2e+05
4e+05
6e+05
8e+05
0/0x0/1
0/1x0/0
0/1x0/1

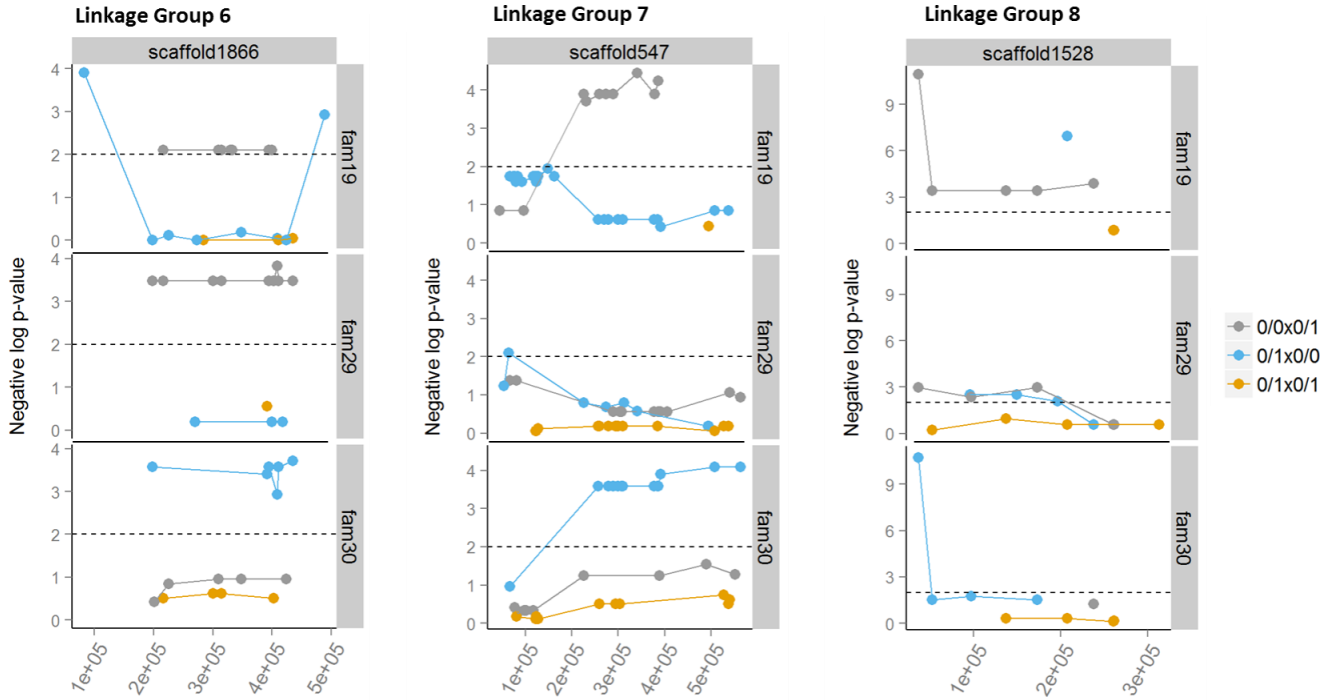
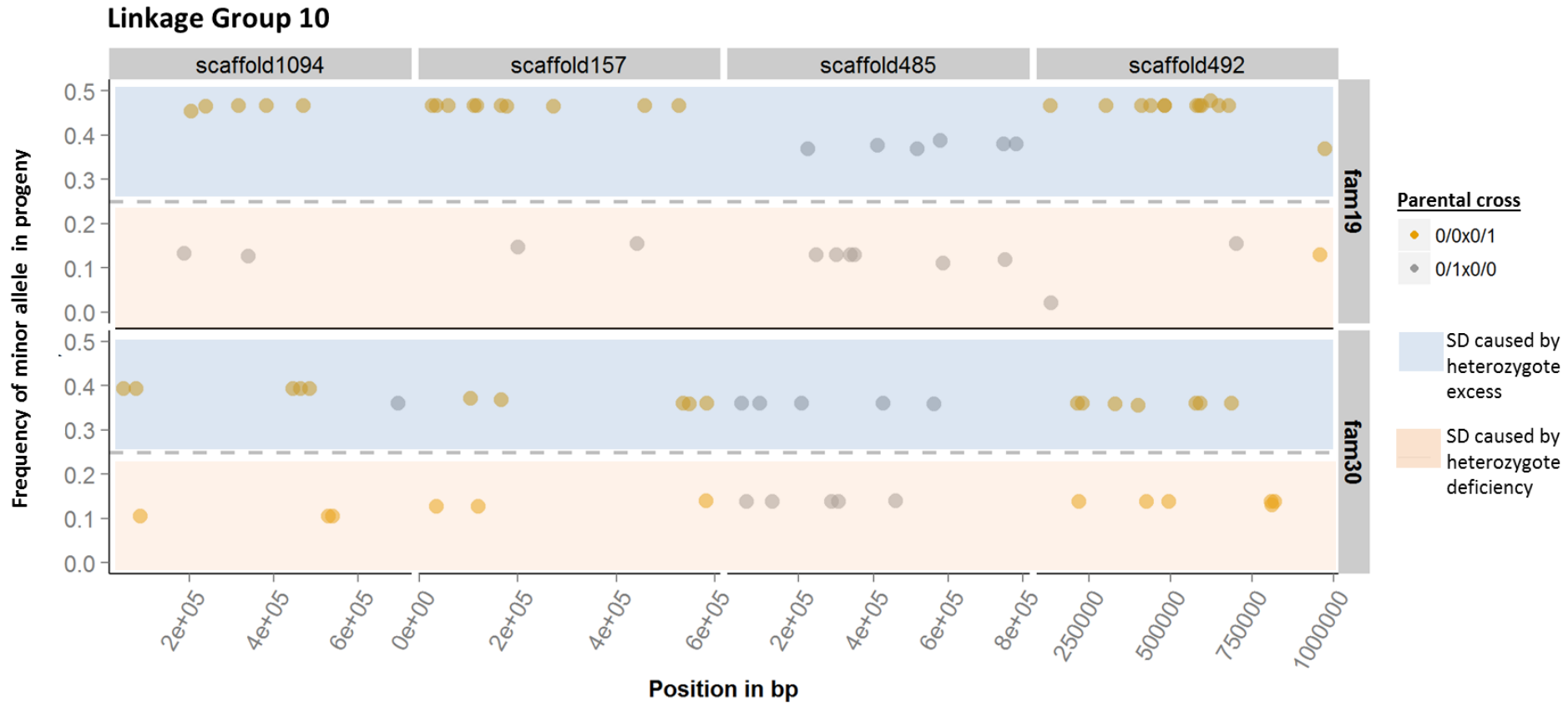


Figure 3. Distribution of SDRs for the twelve selected scaffolds (5 scaffolds for LG 10 + 7 scaffolds per LG) across three Pacific oyster families. Family 19, family 29 and family 30 are coded as fam19, fam29, and fam30, respectively. The x-axes corresponds to the scaffold length in bp, and the y-axes corresponds to the –log10(p-value) of the chi2 goodness of fit test. Dots and lines are colour-coded according to the parental segregation type: grey circles represent SNP markers with a sire x dam "0/0 x 0/1" (or "AA x AB") parental cross; blue circles represent the reciprocal, "0/1 x 0/0" (or "AB x AA") cross; and orange dots represent the segregation results for the "0/1 x 0/1" (or "AB x AB") parental crosses. The vertical black dotted line indicates the negative log value of the corrected significance p-value thresholds.


Linkage Group 10
scaffold1094
scaffold157
scaffold485
scaffold492
fam19
fam30
Frequency of minor allele in progeny
0.5
0.4
0.3
0.2
0.1
0.0
2e+05
4e+05
6e+05
0e+00
2e+05
4e+05
6e+05
2e+05
4e+05
6e+05
8e+05
250000
500000
750000
1000000
Position in bp
Parental cross
0/0x0/1
0/1x0/0
SD caused by heterozygote excess
SD caused by heterozygote deficiency

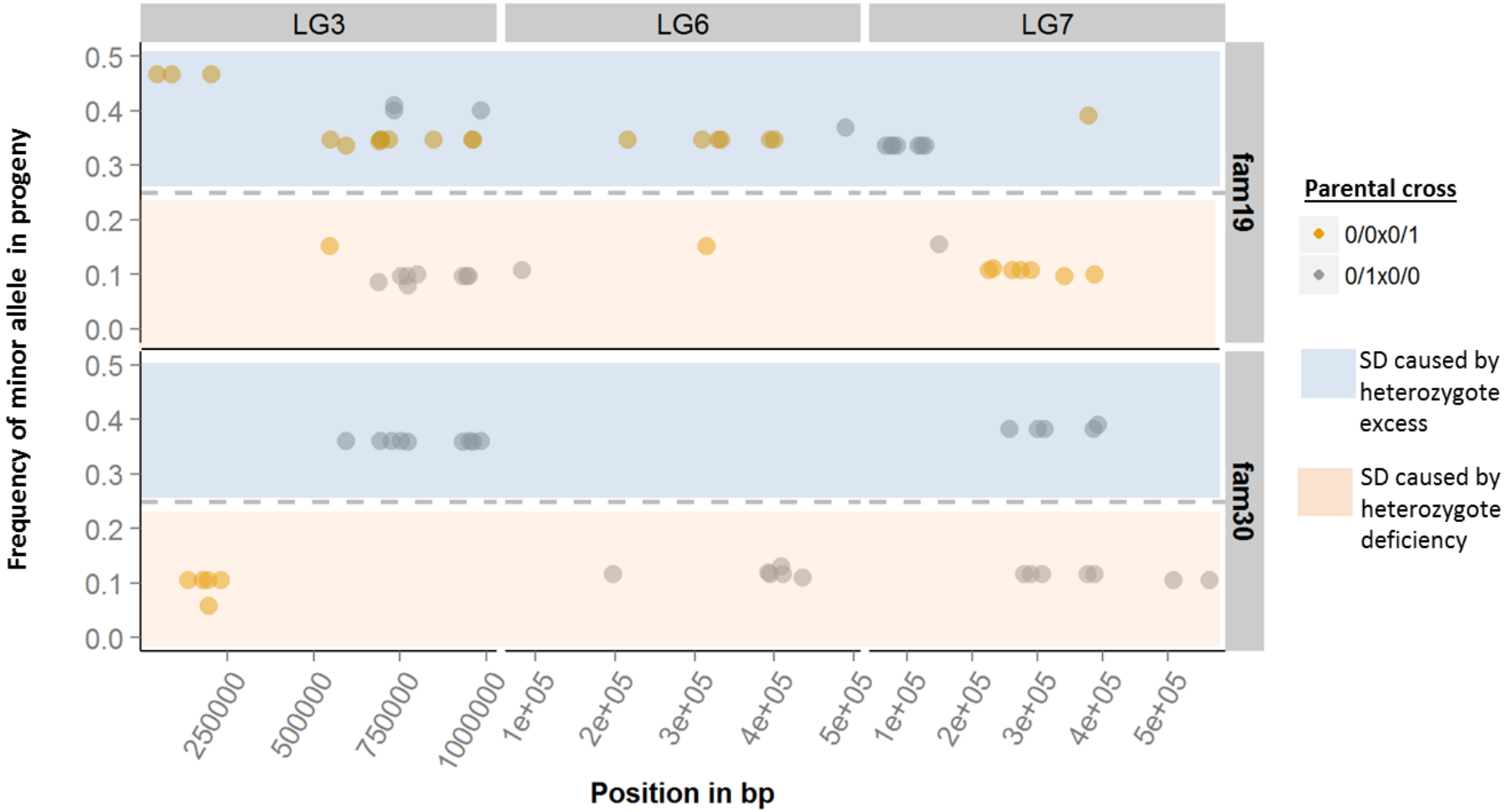

LG3
LG6
LG7
fam19
fam30
Frequency of minor allele in progeny
0.5
0.4
0.3
0.2
0.1
0.0
250000
500000
750000
1000000
1e+05
2e+05
3e+05
4e+05
5e+05
Position in bp
Parental cross
0/0x0/1
0/1x0/0
SD caused by heterozygote excess
SD caused by heterozygote deficiency

Figure 4. Allele frequency of distorted SNP markers across the scaffolds that were 'consistently' distorted in families 19 and 30. Only markers that showed significant distortion after correction for multiple testing are shown. Circles represent the allele frequency of the minor allele in the progeny, which is colored coded according to parental segregation type (sire x dam): "0/0 x 0/1" crosses are represented by orange circles, and "0/1 x 0/0" crosses are represented by a grey circles. The dotted line indicates the expected minor allele frequency (i.e., 0.25) if the marker conformed to Mendelian inheritance. If the minor allele frequency circle is above the dotted line (blue shaded area), then the SD is caused by an excess of heterozygote genotypes. On the contrary, if the minor allele frequency circle is below the dotted line (pink shaded area), the distortion is caused by a deficiency of heterozygote genotypes in the progeny.

To investigate whether the SD was associated with the transmission of the maternal or paternal alleles, we used the high discriminatory power of the RAD-Seq haplotype segregation data. We take advantage of the fact that some haplotype crosses exhibit more than two alleles at the SDRs, and therefore provide a means of testing parent-of-origin effects, as precise tracking of each parental allele is possible. Only two types of informative crosses were detected "AB x CD" and "AB x AC", which represented a low fraction of all the available segregation types (as most crosses were uninformative for the analysis). The results obtained from fitting each parent separately for an allele expectation of a 1:1 ratio are shown in Table 4. The data suggest that both paternal and maternal effects are present, although a higher frequency of distortions driven by the transmission of a female allele is observed (5 out of 6 tests).

Table 4. Estimation of paternal and maternal allelic effects on SDRs identified in different families and Pacific oyster linkage groups (LGs). Chi-square values are given for the paternal and maternal analysis.

Family ID	LG	Mating type	Paternal	Maternal
19	3	AB x CD	9*	6.3*
19	10	AB x AC	3.3	15**
19	10	AB x AC	8.1*	0.6
30	3	AB x AC	0.2	5*
30	3	AB x AC	0.4	5.3*
30	6	AB x AC	1.2	5.8*

***significant (p-value<0.05), ** highly significant (p-value<0.001)**

4.4 Discussion

4.4.1 Performance of RAD-Seq in the Pacific oyster: a comparison of genotyping platforms

In this study, we evaluated the performance of a widely used high-throughput genotyping technique – RAD-Seq – against a recently developed Pacific oyster *C. gigas* SNP array. Our approach was to estimate a platform error rate based on the Mendelian errors identified by a segregation analysis performed in three nuclear families genotyped with the technologies above. We work with the premise that the Pacific oyster genome, similarly to other bivalve genomes (Šatović and Plohl, 2017, Murgarella et al., 2016), show a myriad of genome complexities that represent significant challenges for accurate genotyping, as others have observed (Hedgecock et al., 2004, Lemer et al., 2011, MacAvoy et al., 2008). However, during the design of an array, there is a bias towards selecting SNPs in stable genomic regions with lower levels of polymorphisms, which are generally easier to genotype (Lachance and Tishkoff, 2013). On the contrary, the RAD-Seq method randomly interrogates different regions (Davey and Blaxter, 2010, Miller et al., 2007), which may lead to increased genotyping artifacts in a challenging genome. This fundamental distinction between genotyping platforms allowed us to detect (i) a base-line error, represented by the Mendelian errors of the SNP array, and (ii) a RAD-Seq genotyping error, derived from the comparison of a number of Mendelian errors detected for the SNP array versus the RAD-Seq method. To our knowledge, no other comparative assessment of genome-wide genotyping technologies has been undertaken for a bivalve species.

The results of the segregation analysis indicate that Mendelian errors are abundant when families are genotyped with the RAD-Seq method (41% of the haplotype markers were affected). This degree of error is among the highest described for any platform (Pompanon et al., 2005) and may preclude any reliable population or evolutionary inference based on RAD-Seq-derived markers. The impact, however, would be minor in trait mapping studies, as errors would be easily detected from pedigree information. Among the factors that are most likely to explain this high RAD-Seq genotyping error rate are errors in the pedigree, quality of the DNA samples, sequencing platform and genome-specific characteristics. Each is discussed in the context of this study. First, an error in the pedigree is discarded as an explanatory factor because an identity-by-state clustering analysis based on the SNP array data (~25,000 SNPs) clearly distinguishes the three family groups (Gutierrez et al., 2017), supporting the family integrity of the samples used in this study. The other factor that may

have influenced the high rate of Mendelian errors is DNA quality. High-quality DNA is essential for the RAD-Seq method because it leads to a more consistent and efficient shearing of the DNA fragments (Andrews et al., 2016), which in turn leads to a theoretically unbiased sampling (sequencing) of loci across individuals. In our study, the DNA quality of the offspring samples showed variable degrees of fragmentation (data not shown), which may have affected the adequate sampling of both alleles. However, if due to poor DNA quality we systematically sampled one instead of both alleles (allelic dropout), then we would expect most of the segregation distortions to be caused by an excess of homozygote genotypes in the offspring (or conversely, a heterozygote deficiency). However, the majority of the distortions identified in the RAD-Seq dataset (62%) were caused by a deficiency (and not an excess) of homozygote genotypes (i.e., excess of heterozygotes). Therefore, variable input DNA quality is unlikely to have caused the high error rates. Another explanatory factor could have been the RAD-Seq method itself. As discussed in previous Chapters, RAD-Seq is not exempt from bias, like any other genotyping technique. However, in Chapter 3 of this thesis, and as part of a *de novo* assembly optimization procedure, we sequenced a set of technical replicates from two offspring samples used in the segregation analysis discussed herein. From the Pacific oyster technical replicates, we estimated a genotyping error rate of 3.5% for the RAD-Seq platform, a value that is 11-fold less than our current estimate based on Mendelian errors. The magnitude of this difference suggests that genome-specific characteristics, reflecting a biological factor, are likely explaining why Mendelian errors are so abundant in the RAD-Seq data, and for this matter, also in the SNP array (~8%). It is worth mentioning that other genotyping platforms have been developed for the Pacific oyster, and that these usually show low error rates determined from replicate reactions or cross-technology comparisons (e.g., 0% for SNPs on a Illumina GoldenGate Hedgecock et al. (2015); 0.034% for a high-density 190k SNP array Qi et al. (2017)). Nevertheless, when for the same technologies genotyping accuracy is estimated according to family data, higher errors are reported, such as the 7.4% detected by Qi et al. (2017) for the high-density SNP array. Under the possibility that the pedigree structure is decoupled to some extent from the genotypic data, the default assumption that genotypes are correct if they conform to Mendelian expectations would be inadequate. In the next Chapter, we explore potential biological sources of Mendelian inconsistencies in the Pacific oyster genotyping data.

4.4.2 Segregation distortion: single-marker analysis

After the removal of loci showing Mendelian errors, a high agreement between the segregation analyses of markers genotyped with a SNP-array and the RAD-Seq method was observed. In family 19, 33% of the SNPs from the array were significantly distorted, whereas 35% of the RAD-Seq-derived haplotype markers showed deviations from Mendelian inheritance patterns. In family 29, 12% and 15% of the markers from the SNP array and the RAD-Seq method, respectively, were distorted. In family 30 both platforms had the same proportion of distorted markers, 23%. Segregation distortion in our study was mostly caused, across families and mating types, by a bias against homozygote genotypes in the offspring, in contrast with the widespread observation that distortions in bivalves are caused by heterozygote deficiency (Raymond et al., 1997, Beaumont, 1991, Mallet et al., 1985). This shift in the tendency of distortions could be explained by the fact that the parents used in the study were related to some degree, leading to partial inbreeding in the offspring. Indeed, partially inbred bivalve crosses have been shown to produce substantial deficiencies of homozygous genotypes, which are interpreted as selection against recessive deleterious alleles linked to the markers under study (Plough and Hedgecock, 2011, Launey and Hedgecock, 2001). On the contrary, outbred crosses of marine bivalves appear to be different in the sense that distortions usually manifest as a deficiency of heterozygote genotypes (Plough et al., 2016). The notion that bivalves typically show (a phenomenon of) deficiencies of heterozygote genotypes most probably stem from the fact that, for decades, there were fewer genetic studies on inbred families. Furthermore, most of the literature came from field studies, for which heterozygote deficiency was an established phenomenon (Beaumont, 1991, Foltz, 1986, Raymond et al., 1997, Borsa et al., 1991, Toro and Vergara, 1995, Gosling and Wilkins, 1985, Lassen and Turano, 1978). With the advent of NGS technologies and an increasing interest in the development of genomic resources and linkage maps, information from SDs in marine bivalves with different breeding backgrounds will be available for comparison. Although inbreeding may be a feasible explanation for the observed deficiencies of homozygote genotypes, we cannot test this hypothesis since no pedigree records of the parents were available.

Levels of SD have not yet been explored at the high marker density analyzed here (particularly with the SNP array). Therefore a strict comparison with previous literature is difficult, particularly due to the mixture of the genetic backgrounds (inbred vs. outbred) of

the families that different studies have analysed. For the SNP array markers, proportions of distortions ranged from 12% (1,148 SNPs) to 33% (3,260 SNPs) across families. For the RAD derived haplotype markers, SDs ranged from 15% (140 haplotypes) to 35% (350 haplotypes). In comparison, and with 19 microsatellites genotyped in eight Pacific oyster families, Launey and Hedgecock (2001) reported 21% of non-Mendelian segregation ratios after correction for multiple testing. Reece et al. (2004) found approximately 11% of significant distortions in the oyster *Crassostrea virginica*, also after correction. On the other hand, during the construction of a high-density linkage map for the silver-lipped pearl oyster *Pinctada maxima*, 13% of the SNP markers were shown to be distorted, affecting mainly a single family and specific LGs (Jones et al., 2013). Likewise, in a mapping study on the Pacific oyster, distorted marker appeared in most LGs across five families (Hedgecock et al., 2015). However, a significant difference in the proportion of distorted makers per LG was detected, ranging from 3% in LG1 to 60% in LG 3.

4.4.3 Pattern of SD along genome scaffolds

A reference genome assembly for the Pacific oyster (Zhang et al., 2012) allowed us to co-locate the SNP-chip and the RAD-Seq data. The aim was to detect highly confident SDRs by selecting assemblies (scaffolds) that contained markers from both genotyping platforms that coincide in the presence of significant deviations from Mendelian expectations. Scaffolds containing genotype data were then mapped to a second-generation Pacific oyster linkage map (Hedgecock et al., 2015). To get a representative view of all chromosomes (2n = 20), a single scaffold from each linkage group was sampled; except for LGs 2 and 4, which had no unambiguously mapped scaffolds to examine. In addition, a high-resolution analysis of SDRs was performed for LG 10, as more scaffolds were sampled (n = 5), allowing to examine the chromosome at a higher coverage (~4 Mb).

Because SDRs are represented by clusters of markers closely linked to the gene(s) causing distortion, the highest-skewed genotypic frequencies should occur near the genomic locations of the distorting factor (Lu et al., 2002, Tai et al., 2000). Instead, in this study, patterns of SD on the genome scaffolds were observed in large contiguous blocks, with no tendency of a single marker showing the highest genotypic imbalance. The average length of the scaffolds analyzed averaged ~0.8Mb, and distortions affected from 30 to 92% of the

total scaffold length (Table 3). We found that the extent to which the SDRs overlap was high between two of the three families analyzed, families 19 and 30. This result was unexpected considering that at the level of single markers, it was rare to find SNPs that were commonly distorted in two families, similar to what others have observed (Plough and Hedgecock, 2011). In accordance with previous studies, we observed that markers showing SD tend to cluster in specific LGs (e.g., LG 10) (Jones et al., 2013, Li and Guo, 2004, Hedgecock et al., 2015, McGoldrick and Hedgecock, 1997), indicating that these regions may encompass viability loci. Among the analysed families, SDRs were driven by the female parent in 65% of detected cases (15/23 SDRs detected over three families). This result is confirmed by the segregation analysis of the RAD-Seq-derived haplotype markers at distorted regions. Although both paternal and maternal effects were found to be associated with distortions (Table 4), for most of the crosses analysed (5 out of 6) the female parent deviated significantly from a 1:1 allelic ratio. This result would suggest that SD is mainly caused by the overall reduced representation of individuals carrying a particular female allele. However, this finding does not agree with a study performed in outbred Pacific oyster families, where the maternal effect is only slightly higher than the paternal effects (29 vs. 25) (Plough et al., 2016)). Although in this study the exact genetic factors underlying SDRs cannot completely be resolved, the observation that the segregation ratios show a slight tendency to be biased against homozygote genotypes suggests that zygotic instead of gametic factors are responsible for SDs. Based on previous genetic evidence from Pacific oyster families (Launey and Hedgecock, 2001, Plough and Hedgecock, 2011), it is likely that the distortions detected in this study are caused by post-zygotic viability selection due to the presence of deleterious recessive genes. This high mutational load would also fit with the extreme levels of sequence polymorphisms observed in the present study (one SNP per 48bp). To date, however, no direct evidence for a high mutation rate in any bivalve species has been provided.

4.5 References

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology,* 215**,** 403-410.

ANDREWS, K. R., GOOD, J. M., MILLER, M. R., LUIKART, G. & HOHENLOHE, P. A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics,* 17**,** 81-92.

BAIRD, N. A., ETTER, P. D., ATWOOD, T. S., CURREY, M. C., SHIVER, A. L. & LEWIS, Z. A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One,* 3, e3376.

BARTON, M. D. & BARTON, H. A. 2012. Scaffolder - software for manual genome scaffolding. *Source Code for Biology and Medicine,* 7**,** 4.

BEAUMONT, A. R. 1991. Genetic studies of laboratory reared mussels, Mytilus edulis: heterozygote deficiencies, heterozygosity and growth. *Biological Journal of the Linnean Society,* 44**,** 273-285.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological),* 57**,** 289-300.

BIERNE, N., LAUNEY, S., NACIRI-GRAVEN, Y. & BONHOMME, F. 1998. Early Effect of Inbreeding as Revealed by Microsatellite Analyses on Ostrea edulis Larvae. *Genetics,* 148**,** 1893-1906.

BORSA, P., ZAINURI, M. & DELAY, B. 1991. Heterozygote deficiency and population structure in the bivalve Ruditapes decussatus. *Heredity,* 66**,** 1-8.

CATCHEN, J., HOHENLOHE, P. A., BASSHAM, S., AMORES, A. & CRESKO, W. A. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology,* 22**,** 3124-3140.

CATCHEN, J. M., AMORES, A., HOHENLOHE, P., CRESKO, W. & POSTLETHWAIT, J. H. 2011. Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes|Genomes|Genetics,* 1**,** 171-182.

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T., SHERRY, S. T., MCVEAN, G. & DURBIN, R. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-2158.

DAVEY, J. L. & BLAXTER, M. W. 2010. RAD-seq: Next-generation population genetics. *Briefings in Functional Genomics*, 9, 416-423.

DUARTE, C. M., MARBÁ, N. & HOLMER, M. 2007. Rapid Domestication of Marine Species. *Science,* 316**,** 382-383.

ENGELSTÄDTER, J. & HURST, G. D. D. 2009. The Ecology and Evolution of Microbes that Manipulate Host Reproduction. *Annual Review of Ecology, Evolution, and Systematics,* 40**,** 127-149.

FALCONER, D. S. 1975. *Introduction to quantitative genetics*, Pearson Education India.

FOLTZ, D. W. 1986. Null alleles as a possible cause of heterozygote deficiencies in the oyster Crassostrea virginica and other bivalves. *Evolution,* 40**,** 869-870.

FU, L., NIU, B., ZHU, Z., WU, S. & LI, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics,* 28**,** 3150-3152.

GOSLING, E. M. & WILKINS, N. P. 1985. Genetics of settling cohorts of Mytilus edulis (L.): Preliminary observations. *Aquaculture,* 44**,** 115-123.

GUO, X., HE, Y., ZHANG, L., LELONG, C. & JOUAUX, A. 2015. Immune and stress responses in oysters with insights on adaptation. *Fish & Shellfish Immunology,* 46**,** 107-119.

GUTIERREZ, A. P., TURNER, F., GHARBI, K., TALBOT, R., LOWE, N. R., PEÑALOZA, C., MCCULLOUGH, M., PRODÖHL, P. A., BEAN, T. P. & HOUSTON, R. D. 2017.

Development of a Medium Density Combined-Species SNP Array for Pacific and European Oysters (Crassostrea gigas and Ostrea edulis). *G3: Genes|Genomes|Genetics*, 7, 2209-2218

HARRANG, E., LAPEGUE, S., MORGA, B. & BIERNE, N. 2013. A High Load of Non-neutral Amino-Acid Polymorphisms Explains High Protein Diversity Despite Moderate Effective Population Size in a Marine Bivalve With Sweepstakes Reproduction. *G3: Genes|Genomes|Genetics,* 3**,** 333-341.

HEDGECOCK, D., GAFFNEY, P. M., GOULLETQUER, P., GUO, X., REECE, K. & WARR, G. W. 2005. THE CASE FOR SEQUENCING THE PACIFIC OYSTER GENOME. *Journal of Shellfish Research,* 24**,** 429-441.

HEDGECOCK, D., LI, G., HUBERT, S., BUCKLIN, K. & RIBES, V. 2004. Widespread null alleles and poor cross-species amplification of microsatellite DNA loci cloned from the Pacific oyster, Crassostrea gigas. *Journal of shellfish research,* 23**,** 379-385.

HEDGECOCK, D., SHIN, G., GRACEY, A. Y., DEN BERG, D. V. & SAMANTA, M. P. 2015. Second-Generation Linkage Maps for the Pacific Oyster Crassostrea gigas Reveal Errors in Assembly of Genome Scaffolds. *G3: Genes|Genomes|Genetics,* 5**,** 2007-2019.

JONES, D. B., JERRY, D. R., KHATKAR, M. S., RAADSMA, H. W. & ZENGER, K. R. 2013. A high-density SNP genetic linkage map for the silver-lipped pearl oyster, Pinctada maxima: a valuable resource for gene localisation and marker-assisted selection. *BMC Genomics,* 14**,** 810.

KELLEY, J. L., BROWN, A. P., THERKILDSEN, N. O. & FOOTE, A. D. 2016. The life aquatic: advances in marine vertebrate genomics. *Nature Reviews Genetics,* 17**,** 523-534.

KULMUNI, J., SEIFERT, B. & PAMILO, P. 2010. Segregation distortion causes large-scale differences between male and female genomes in hybrid ants. *Proceedings of the National Academy of Sciences of the United States of America,* 107**,** 7371-7376.

LACHANCE, J. & TISHKOFF, S. A. 2013. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *BioEssays : news and reviews in molecular, cellular and developmental biology,* 35**,** 780-786.

LASSEN, H. H. & TURANO, F. J. 1978. Clinal variation and heterozygote deficit at the lap-locus in Mytilus edulis. *Marine Biology,* 49**,** 245-254.

LAUNEY, S. & HEDGECOCK, D. 2001. High genetic load in the Pacific oyster Crassostrea gigas. *Genetics,* 159**,** 255-265.

LEMER, S., ROCHEL, E. & PLANES, S. 2011. Correction Method for Null Alleles in Species with Variable Microsatellite Flanking Regions, A Case Study of the Black-Lipped Pearl Oyster Pinctada margaritifera. *Journal of Heredity,* 102**,** 243-246.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* 25**,** 1754-1760.

LI, W. & GODZIK, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 22**,** 1658-1659.

LUO, L. & XU, S. 2003. Mapping viability loci using molecular markers. *Heredity (Edinb),* 90**,** 459-467.

LYTTLE, T. W. 1993. Cheaters sometimes prosper: distortion of mendelian segregation by meiotic drive. *Trends in Genetics,* 9**,** 205-210.

MACAVOY, E. S., WOOD, A. R. & GARDNER, J. P. A. 2008. Development and evaluation of microsatellite markers for identification of individual Greenshell™ mussels (Perna canaliculus) in a selective breeding programme. *Aquaculture,* 274**,** 41-48.

MALLET, A. L., ZOUROS, E., GARTNER-KEPKAY, K. E., FREEMAN, K. R. & DICKIE, L. M. 1985. Larval viability and heterozygote deficiency in populations of marine bivalves: evidence from pair matings of mussels. *Marine Biology,* 87**,** 165-172.

MCGOLDRICK, D. J. & HEDGECOCK, D. 1997. Fixation, Segregation and Linkage of Allozyme Loci in Inbred Families of the Pacific Oyster Crassostrea gigas (Thunberg): Implications for the Causes of Inbreeding Depression. *Genetics,* 146**,** 321-334.

MILLER, M. R., DUNHAM, J. P., AMORES, A., CRESKO, W. A. & JOHNSON, E. A. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research,* 17**,** 240-248.

MURGARELLA, M., PUIU, D., NOVOA, B., FIGUERAS, A., POSADA, D. & CANCHAYA, C. 2016. A First Insight into the Genome of the Filter-Feeder Mussel Mytilus galloprovincialis. *PLOS ONE,* 11**,** e0151561.

PLOUGH, L. V. 2016. Genetic load in marine animals: a review. *Current Zoology,* 62**,** 567-579.

PLOUGH, L. V. & HEDGECOCK, D. 2011. Quantitative Trait Locus Analysis of Stage-Specific Inbreeding Depression in the Pacific Oyster Crassostrea gigas. *Genetics,* 189**,** 1473-1486.

PLOUGH, L. V., SHIN, G. & HEDGECOCK, D. 2016. Genetic inviability is a major driver of type III survivorship in experimental families of a highly fecund marine bivalve. *Molecular Ecology,* 25**,** 895-910.

POMPANON, F., BONIN, A., BELLEMAIN, E. & TABERLET, P. 2005. Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics,* 6**,** 847-846.

QI, H., SONG, K., LI, C., WANG, W., LI, B., LI, L. & ZHANG, G. 2017. Construction and evaluation of a high-density SNP array for the Pacific oyster (Crassostrea gigas). *PLOS ONE,* 12**,** e0174007.

RAYMOND, M., XE, XE, NT, XF, L., R., THOMAS, F., XE, DERIC, ROUSSET, F., XE, OIS, DE, M., XFC, S, T., RENAUD, F., XE & OIS 1997. Heterozygote deficiency in the mussel Mytilus edulis species complex revisited. *Marine Ecology Progress Series,* 156**,** 225-237.

REECE, K. S., RIBEIRO, W. L., GAFFNEY, P. M., CARNEGIE, R. B. & ALLEN, J. S. K. 2004. Microsatellite Marker Development and Analysis in the Eastern Oyster (Crassostrea virginica): Confirmation of Null Alleles and Non-Mendelian Segregation Ratios. *Journal of Heredity,* 95**,** 346-352.

ROBLEDO, D., PALAIOKOSTAS, C., BARGELLONI, L., MARTÍNEZ, P. & HOUSTON, R. 2017. Applications of genotyping by sequencing in aquaculture breeding and genetics. *Reviews in Aquaculture***,** 1-13.

RUTKOWSKA, J. & BADYAEV, A. V. 2008. Meiotic drive and sex determination: molecular and cytological mechanisms of sex ratio adjustment in birds. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 363**,** 1675-1686.

ŠATOVIĆ, E. & PLOHL, M. 2017. Two new miniature inverted-repeat transposable elements in the genome of the clam Donax trunculus. *Genetica*, 145, 379-385.

SAUPE, S. J. 2012. A fungal gene reinforces Mendel's laws by counteracting genetic cheating. *Proceedings of the National Academy of Sciences,* 109**,** 11900-11901.

SAUVAGE, C., BIERNE, N., LAPÈGUE, S. & BOUDRY, P. 2007. Single Nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster Crassostrea gigas. *Gene,* 406**,** 13-22.

SEYMOUR, D. K., CHAE, E., ARIÖZ, B. I., KOENIG, D. & WEIGEL, D. 2017. The genetic architecture of recurrent segregation distortion in Arabidopsis thaliana. *bioRxiv*.

TAO, Y., HARTL, D. L. & LAURIE, C. C. 2001. Sex-ratio segregation distortion associated with reproductive isolation in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America,* 98**,** 13183-13188.

TAYLOR, D. & INGVARSSON, P. 2003. Common Features of Segregation Distortion in Plants and Animals. *Genetica,* 117**,** 27-35.

TORO, J. E. & VERGARA, A. M. 1995. Evidence for Selection Against Heterozygotes: Post-Settlement Excess of Allozyme Homozygosity in a Cohort of the Chilean Oyster, Ostrea chilensis Philippi, 1845. *The Biological Bulletin,* 188**,** 117-119.

WANG, L., SONG, L., CHANG, Y., XU, W., NI, D. & GUO, X. 2005. A preliminary genetic map of Zhikong scallop (Chlamys farreri Jones et Preston 1904). *Aquaculture Research,* 36**,** 643-653.

WILLIS, S. C., HOLLENBECK, C. M., PURITZ, J. B., GOLD, J. R. & PORTNOY, D. S. 2017. Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. *Molecular Ecology Resources***,** 17, 955-965.

YATES, A., AKANNI, W., AMODE, M. R., BARRELL, D., BILLIS, K., CARVALHO-SILVA, D., CUMMINS, C., CLAPHAM, P., FITZGERALD, S., GIL, L., GIRÓN, C. G., GORDON, L., HOURLIER, T., HUNT, S. E., JANACEK, S. H., JOHNSON, N., JUETTEMANN, T., KEENAN, S., LAVIDAS, I., MARTIN, F. J., MAUREL, T., MCLAREN, W., MURPHY, D. N., NAG, R., NUHN, M., PARKER, A., PATRICIO, M., PIGNATELLI, M., RAHTZ, M., RIAT, H. S., SHEPPARD, D., TAYLOR, K., THORMANN, A., VULLO, A., WILDER, S. P., ZADISSA, A., BIRNEY, E., HARROW, J., MUFFATO, M., PERRY, E., RUFFIER, M., SPUDICH, G., TREVANION, S. J., CUNNINGHAM, F., AKEN, B. L., ZERBINO, D. R. & FLICEK, P. 2016. Ensembl 2016. *Nucleic Acids Research,* 44**,** D710-D716.

ZHANG, G., FANG, X., GUO, X., LI, L., LUO, R., XU, F., YANG, P., ZHANG, L., WANG, X., QI, H., XIONG, Z., QUE, H., XIE, Y., HOLLAND, P. W. H., PAPS, J., ZHU, Y., WU, F., CHEN, Y., WANG, J., PENG, C., MENG, J., YANG, L., LIU, J., WEN, B., ZHANG, N., HUANG, Z., ZHU, Q., FENG, Y., MOUNT, A., HEDGECOCK, D., XU, Z., LIU, Y., DOMAZET-LOSO, T., DU, Y., SUN, X., ZHANG, S., LIU, B., CHENG, P., JIANG, X., LI, J., FAN, D., WANG, W., FU, W., WANG, T., WANG, B., ZHANG, J., PENG, Z., LI, Y., LI, N., WANG, J., CHEN, M., HE, Y., TAN, F., SONG, X., ZHENG, Q., HUANG, R., YANG, H., DU, X., CHEN, L., YANG, M., GAFFNEY, P. M., WANG, S., LUO, L., SHE, Z., MING, Y., HUANG, W., ZHANG, S., HUANG, B., ZHANG, Y., QU, T., NI, P., MIAO, G., WANG, J., WANG, Q., STEINBERG, C. E. W., WANG, H., LI, N., QIAN, L., ZHANG, G., LI, Y., YANG, H., LIU, X., WANG, J., YIN, Y. & WANG, J. 2012b. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature,* 490**,** 49-54.

ZHANG, L., LI, L., GUO, X., LITMAN, G. W., DISHAW, L. J. & ZHANG, G. 2015. Massive expansion and functional divergence of innate immune genes in a protostome. *Scientific Reports,* 5**,** 8693.

ZHOU, W., TANG, Z., HOU, J., HU, N. & YIN, T. 2015. Genetic Map Construction and Detection of Genetic Loci Underlying Segregation Distortion in an Intraspecific Cross of Populus deltoides. *PLOS ONE,* 10**,** e0126077.

CHAPTER 5

Using locus sequence coverage to investigate null alleles in Pacific oyster RAD-Seq data

5.1 Introduction

Null alleles at genetic markers are alleles that fail to be detected. They can be the result of a deletion encompassing the marker (Crooks et al., 2013) or polymorphisms at the binding site of PCR primers, which is relevant for microsatellite analysis (Broquet and Petit, 2004, Dakin and Avise, 2004). In the case of genotyping by sequencing approaches, polymorphisms occurring at restriction enzyme recognition sites can be a common source of null alleles (Andrews et al., 2016, Cooke et al., 2016). Due to the fact that a locus that is heterozygous for a null allele (e.g., "AØ") will be genotyped as a homozygote (i.e., "AA"), the segregation of null alleles within a family/pedigree structure may result in apparent inheritance incompatibilities between parent and offspring genotypes. As such, null alleles may be problematic for genetic analysis as they could lead to false parentage exclusion (Dakin and Avise, 2004), or potentially artificially inflate population structure in cases of significant underlying population differentiation (Chapuis and Estoup, 2007). However, instead of discarding markers carrying a null allele some researchers have used them as a means of discovering structural genomic variation in humans (McCarroll et al., 2008, Conrad et al., 2006, Kohler and Cutler, 2007), chickens (Crooks et al., 2013) and cattle (Seroussi et al., 2010). This exemplifies the utility of understanding the underlying factors of genotypic incompatibilities in pedigree data, as it may provide an opportunity for an improved view of the polymorphism landscape of a species.

Null alleles are ubiquitous in genetic studies of bivalve species. Based on the observation of unexpected genotypes in the offspring of pair crosses, null alleles in marine bivalves affect from 16% to 51% of the employed markers (Reece et al., 2004, Launey and Hedgecock, 2001, McGoldrick et al., 2000, Hedgecock et al., 2004). A few studies have empirically confirmed the wide-spread nature of null alleles in marine bivalves and their fundamental impact on biological interpretation. In a striking example in the Black-Lipped oyster *Pinctada margaritifera*, Lemer et al. (2011) completely corrected deviations from Hardy Weinberg equilibrium (HWE) by re-designing the PCR primers for microsatellite loci. Another study in the American oyster *Crassostrea virginica* (Hare et al., 1996) revealed that the genotype scoring of some loci appeared to vary as a function of varying PCR conditions, suggesting a high risk of erroneous scoring of genotypes; for example, a locus that showed departures from HWE in 58% of the populations examined, conformed to HWE after primer redesign. These studies highlight the pervasiveness of technical biases in PCR assays in

bivalves, and of particular concern, the difficulty of detecting them in natural populations, given the circular argument of using agreement with HWE as a criterion for assessing the validity of an assay. The most reliable approach to evaluate genotype errors is direct verification of Mendelian transmission through pedigree analysis. In the search for developing parentage assignment tools, researchers have encountered several technical issues (Pino-Querido et al., 2015, Morvezen et al., 2013, Reece et al., 2004, MacAvoy et al., 2008) that, although not precluding parental testing, confirm that bivalve species display considerable locus complexity. For instance, three of these studies (MacAvoy et al., 2008, Reece et al., 2004, Zhan et al., 2007) sequenced the flanking region of a subset of microsatellite loci that were being evaluated for parental testing, as means of ensuring their allelic state. A substantial number of polymorphisms in the flanking region were found at most loci, indicating that genetic variation not only occurred at the microsatellite repeat motif itself but also in the surrounding DNA sequence. Due to the high sequence polymorphisms of marine bivalves (Zhang et al., 2012a, Harrang et al., 2013, Sauvage et al., 2007, Pallavicini et al., 2008), a high incidence of technical artefacts may be expected, which may severely impact the development and application of markers, along with the subsequent interpretation of genetic data. Furthermore, the failure to incorporate the interaction between genetic assays and genome complexities during the stages of experimental design and data interpretation has led to a few occasions in which conclusions were shown to be completely biased by technical artifacts, e.g., Ren et al. (2009) correcting Yu et al. (2008). Considering the high risk of bias in genetic data obtained from species with complex and poorly understood genomes, of particular relevance is the examination of potential sources of technical errors in technologies that are currently being used for high-throughput genotyping of non-model organisms.

Restriction site associated (RAD) sequencing is a cost-effective means of generating large numbers of SNPs in virtually any organism. Several potential sources of bias have been described for the RAD-Seq protocols (Andrews et al., 2016). A relevant source of bias is the presence of polymorphisms in restriction enzyme recognition sites, which may create false homozygous genotype calls due to the presence of a null allele (Davey et al., 2013, Arnold et al., 2013, Ilut et al., 2014, Gautier et al., 2013). In a bivalve genome, where rates of sequence and structural polymorphisms are high compared to most other species (Rico et al., 2017, Zhang et al., 2012b), a greater frequency of null alleles is expected. In the face of the high number of inheritance incompatibilities detected in Chapter 4 of this thesis, it is

reasonable to suggest that null alleles may be explaining at least a proportion of these Mendelian errors. Additionally, the fact that we have observed a bimodal distribution in the histograms of coverage per locus for all individuals and species analyzed in this thesis (example in Chapter 3; Figure 6) may indicate that null alleles are segregating at significant magnitudes in the bivalve RAD-Seq data (i.e., the peak of lower coverage may reflect the failure to detect one of the alleles at a given locus, or genuine hemizygosity at the genomic region). Evidence for the hypothesis of high levels of null alleles in bivalve genetic data has support in data obtained from the (Pacific) Oyster Genome Consortium that indicated that 42% of microsatellite loci occurred in the hemizygous state due to indel variation (personal communication in Rico et al. (2017)).

The aim of this Chapter was to test the hypothesis that the extensive Mendelian errors detected in the Pacific oyster RAD-Seq data (Chapter 4) are caused by null alleles. Our approach was to establish an association between markers (i) showing Mendelian errors and (ii) their patterns in parents and offspring when coverage is taken into account in addition to the simple genotype call information. The rationale is that haplotypes erroneously typed as homozygous due to a putative null allele (i.e., "AØ") may exhibit lower coverage on average – specifically that they should come from the 'low-coverage' peak of the individual's coverage per locus bimodal distribution. Following the same logic, markers that show patterns that fit with the presence of a null allele (i.e., presence of genotypes corresponding to low-coverage homozygotes) should always be flagged as a locus with high potential for Mendelian incompatibility. By evaluating how many violations of Mendelian inheritance are consistent with the presence of null alleles, we aim to derive an estimate of their abundance in the Pacific oyster RAD-Seq data and their potential contribution to the highly unorthodox segregation patterns typically observed in bivalve nuclear families.

5.2 Material and Methods

5.2.1 Animals and Sequence Data

RAD-Seq data from the three oyster families (total n = 68) described in Chapter 4 were analyzed. In Chapter 4, genotypes were called separately within each family because a preliminary analysis in Stacks v1.40 revealed that the number of loci in a catalogue built from cross-referencing the family-specific loci was lower than a family-based genotype

calling approach (Appendix, Section1: Figure 1). However, here we aim to take advantage of the fact that families 29 and 30 share the same dam, as it allows us to test if null alleles of maternal origin are being transmitted to the progeny. If at the same locus we observe that offspring from families 29 and 30 show a segregation pattern consistent with the inheritance of a maternal null allele, then the Mendelian incompatibility is attributed to a null allele instead of a potential genotyping error.

After quality filtering and trimming, reads from individual oysters belonging to all three families were assembled and genotyped *de novo* using Stacks v1.4 (Catchen et al., 2013, Catchen et al., 2011). A minimum stack depth of 3 and a maximum of 6 mismatches were allowed to build a locus within an individual. In addition, a gapped alignment was enabled to allow for the detection of indels. A catalogue of representative loci was obtained for each individual by merging loci using a maximum sequence mismatch of 8. After samples are matched back to the family loci catalogue, we exported the genotype data as RAD loci haplotypes only if the marker was present in 70% of the individuals at a minimum coverage of 8 reads.

Haplotype markers genotyped across the three oyster families were then categorized into one of the following haplotype marker status: (i) Mendelian errors, if the genotypes in the parents and offspring showed incompatibilities and therefore did not fit with expected Mendelian inheritance, (ii) a distorted locus (SD), if the marker conformed to expected Mendelian inheritance but offspring genotypes deviated from expected genotype ratios under Mendelian laws, and (iii) normal, if the locus showed a normal inheritance pattern and showed no significant deviation from expected ratios. For the segregation analysis of the loci free of Mendelian errors (i.e., those not in the Mendelian error category) we tested for significant departures from Mendelian expectations within families using a chi-square test ($\alpha = 0.05$).

5.2.2 Detection of null alleles

All individual haplotypes were recoded according to two variables: (i) genotypes status – i.e., either a homozygote or heterozygote haplotype; and (ii) haplotype coverage, either above or below a predefined threshold 37x. This threshold value was determined by estimating approximately half the distance between the two peaks of the bimodal coverage

per locus distribution observed for the oyster samples after read normalization; peaks occurred at ~25x and ~49x. The experimental threshold of 37 (at the theoretical minimum between the two coverage peaks) was chosen to distinguish between the loci with one observed gene copy (putative null alleles) and the putative normal loci with two gene copies (Figure 1). This was taken as a pragmatic approach, expecting a proportion of false positives and false negatives in each category due to the overlap of the coverage distributions.

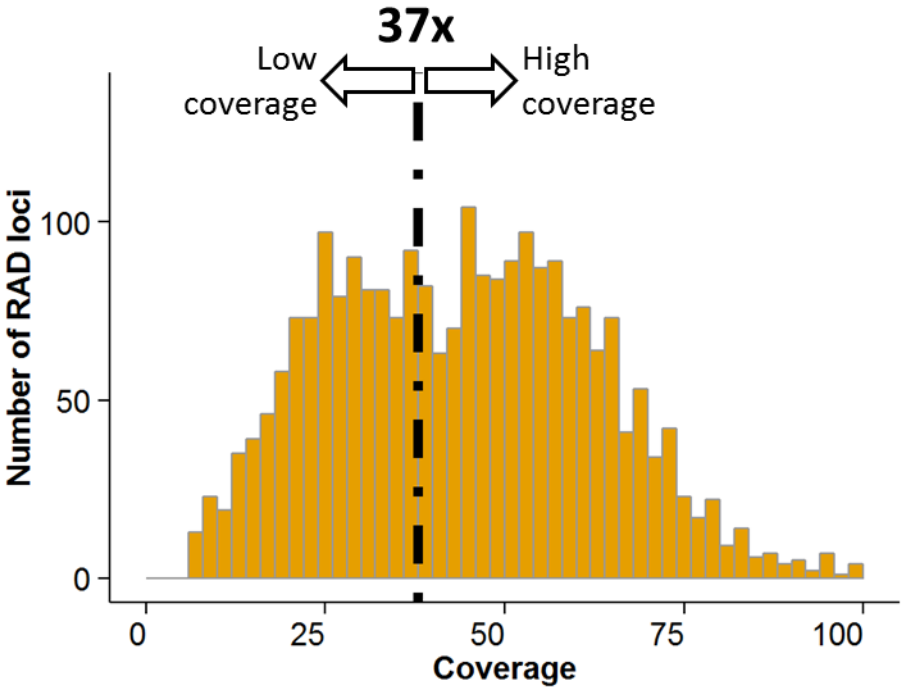


Figure 1. The experimental threshold used to define high and low coverage regions from the histogram of coverage per locus. The y-axis shows the frequency of RAD-loci, and the x-axis shows the read coverage for a given locus.

Following the classification described above, all typed haplotypes were assigned to one of the following bivariate categories (henceforth a 'genotype-coverage' category):

i. homozygote of low coverage (Hom-LC),
ii. homozygote of high coverage (Hom-HC),
iii. heterozygote of low coverage (Het-LC),
iv. and heterozygote of high coverage (Het-HC).

5.2.3 Frequency of 'genotype-coverage' categories in an individual

As a first exploration of the genotype coverage data, we calculated the number of times each 'genotype-coverage' category is present in an individual oyster. To allow for a comparative analysis, each category is expressed as a percentage of the total number of non-missing haplotypes per individual.

5.2.4 Association between genotype-coverage patterns and Mendelian inconsistencies

The presence of null alleles in our dataset was assessed by evaluating the association between Mendelian errors and 'genotype-coverage' patterns derived from parents and offspring. Since we were interested in evaluating the presence of null alleles, both the heterozygotes showing HC and LC are merged into a single heterozygote category. This was done because the relevance of coverage (either HC or LC) is by default minimal in a heterozygous genotype, as both alleles are present and have been detected. Consequently, only three categories were considered when extracting parental and offspring 'genotype-coverage' patterns: homozygote-LC, homozygote-HC, and heterozygotes. For each haplotype marker within a family, a 'genotype-coverage' pattern is defined by taking the frequencies of the abovementioned categories and observing whether a category is present or absent in the (i) parents (parental pattern); and (ii) offspring (offspring pattern). In other words, 'genotype-coverage' data is being represented for the parents and offspring using a binary variable (present/absent). Contingency tables summarizing the joint frequency distribution of 'genotype-coverage' patterns and haplotype status (either Mendelian error, SD or normal) were visualized in mosaic plots using R (R Core Team, 2013). The association between 'genotype-coverage' patterns and proportions of Mendelian errors was assessed

using a Pearson chi^2 test of independence with a p-value threshold of 0.05, after applying a Bonferroni correction for multiple testing.

5.3 Results

A total of 2,830 haplotype markers were identified across the three oyster families. Each marker was categorized by their haplotype status, either as showing (i) Mendelian error, (ii) segregation distortion (SD), or (iii) normal Mendelian inheritance ratios (i.e., no SD) (Table 1).

Table 1. Summary of haplotype marker status. For each oyster family, the frequency of markers showing Mendelian errors, segregation distortion (SD), and normal inheritance are presented. In the last column, the number of informative haplotypes per family is shown, with the percentage from the total amount of markers identified in each family in parenthesis.

Family ID	Mendelian error	SD	Normal inheritance, no SD	Numb. informative haplotypes
Family 19	313	695	326	1,334 (47%)
Family 29	399	746	367	1,512 (53%)
Family 30	387	777	390	1,544 (55%)

5.3.1 Detection of null alleles

For each individual oyster, all markers were categorized according to two variables: haplotype genotype, either homozygous or heterozygous; and coverage, either above or below the pre-defined read depth threshold of 37x. The percentages of each bivariate category (four in total: Hom-HC, Hom-LC, Het-HC and Het-LC) were calculated for each individual, without taking into account missing genotypes.

Within an individual, irrespective of the family, ~47% of the haplotype markers were called as heterozygous. The proportion of homozygous haplotype markers per individual is marginally higher than heterozygous haplotypes, which can be observed in Figure 2. At an

individual level, genome-wide heterozygosity was relatively constant across family members. Notably, within each genotype category (i.e., homozygote versus heterozygote) the ratio between LC and HC genotypes shows a statistically significant difference. Regarding the homozygote genotypes category, the ratio of LC to HC across individuals is around 1.14, whereas for the heterozygote category the ratio is 0.38. Thus, the proportion of heterozygous-LC haplotypes is significantly lower than the equivalent homozygous category (i.e., homozygous-LC) in all three families (chi^2 goodness of fit test: p-value<0.001 in all families). This means that homozygous haplotypes are equally likely to be designated as originating from the low coverage distribution than from the high distribution. On the contrary, the vast majority of heterozygous haplotypes genotyped within individual oysters are found in the 'high coverage' spectrum of the bimodal distribution observed for the coverage per locus plots (Figure 2).

The incidence of null alleles was investigated by evaluating the association between (i) 'genotype-coverage' patterns derived for both parent and offspring genotype data and (ii) the number of Mendelian errors. The rationale behind this was that patterns that fit with the presence of a null allele (i.e., patterns that contain the homozygous-LC) should be associated with a higher frequency of Mendelian errors. A summary of the frequency of the parent and offspring 'genotype-coverage' patterns for each oyster family are shown in Tables 2 and 3, respectively.

Regarding the parental 'genotype-coverage' patterns, the relative frequency of the patterns is consistent across families (Table 2). The most frequent pattern is the "Het x Het" cross, which is uninformative for the detection of null alleles. The following patterns, "Het x Hom-HC" and "Het x Hom-LC", show no statistically significant difference in numbers (chi^2 goodness of fit: p-value = 0.3). For the parental 'genotype-coverage' pattern that is informative for the detection of null alleles (i.e., the "Het x Hom-LC" cross), the sire was equally likely than the dam to be the Hom-LC category (chi^2 goodness of fit: p-value = 0.2). Similar to the parental 'genotype-coverage' patterns, the frequency of the offspring patterns is similar across families (Table 3). The most frequent pattern observed in the offspring is of instances in which a haplotype marker shows three 'genotype-coverage' categories in the offspring: Het, Hom-HC, and Hom-LC.

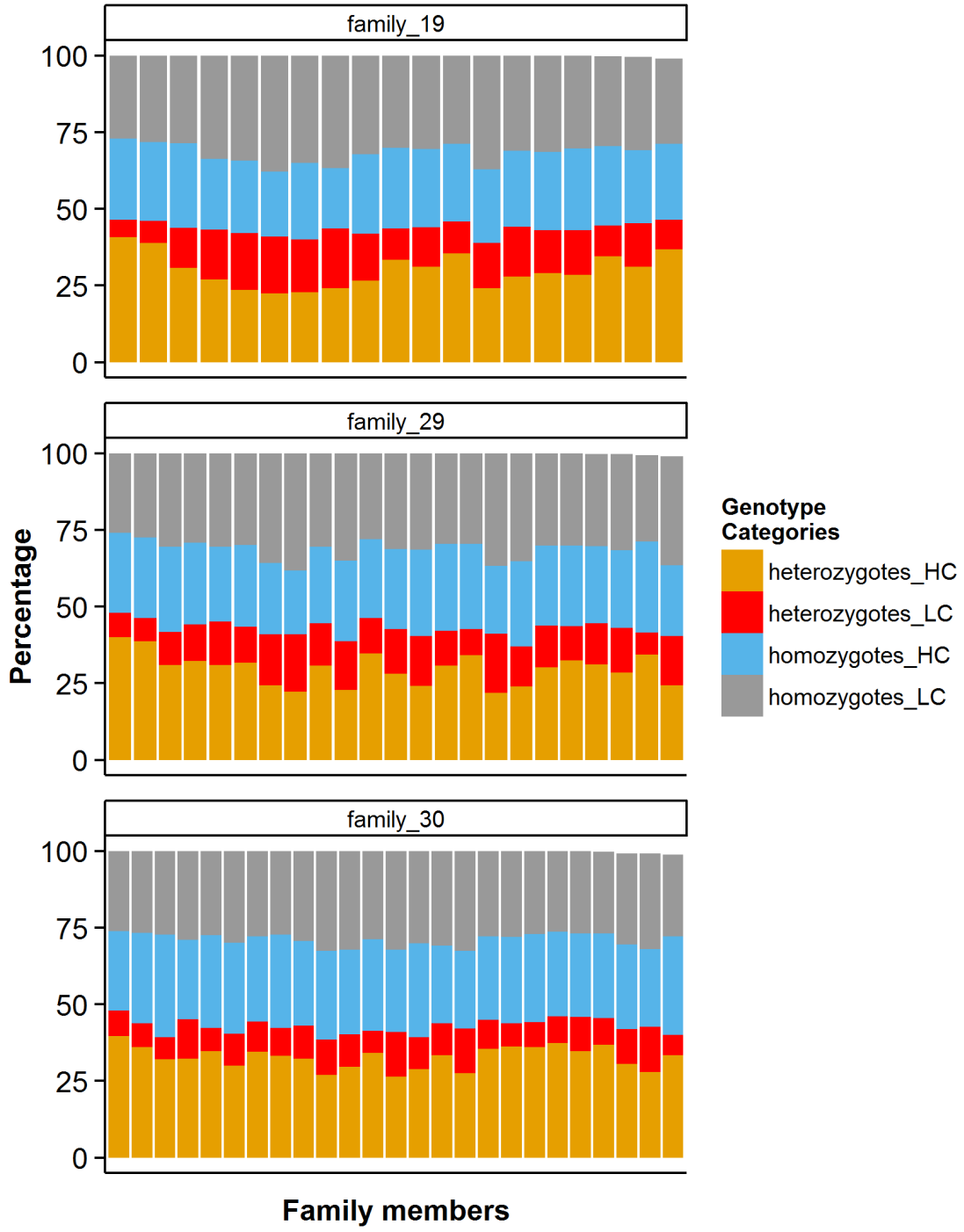


Figure 2. The proportion of Het_LC category is comparatively inferior to other 'genotype-coverage' combinations. In the x-axes the different individuals from a family (parents + offspring) is shown and each bar represents an individual. In the y-axes the percentage of haplotypes belonging to each category is shown and color-coded according to the legend at the right after removal of missing genotypes.

Table 2. Frequency of parental 'genotype-coverage' patterns across families. The bottom row shows the number of informative (haplotype) crosses, with percentages from the total number of markers identified in each family in parenthesis.

Parental pattern	Family 19	Family 29	Family 30
Het x Hom-LC	328	393	398
Het x Hom-HC	419	453	489
Het x Het	442	493	502
Hom-LC x Hom-HC	35	53	43
Hom-HC x Hom-HC	5	8	13
Hom-LC x Hom-LC	105	112	109
Total	1,334 (47%)	1,512 (53%)	1,544 (55%)

Table 3. Frequency of offspring 'genotype-coverage' patterns across families. The bottom row shows the number of informative (haplotype) crosses, with percentages from the total number of markers identified in each family in parenthesis.

Offspring pattern	N^{o} categories	Family 19	Family 29	Family 30
Het | Hom-LC	2	331	325	254
Het |Hom-HC	2	305	351	433
Het	1	51	89	101
Hom-LC | Hom-HC	2	10	10	10
Hom-LC	1	6	2	0
Hom-LC | Het | Hom-HC	3	631	735	756
Total		1,334 (47%)	1,512 (53%)	1,544 (55%)

To detect the segregation of null alleles in our data we explored the relationship between two categorical variables. One variable describes a 'genotype-coverage' pattern, which is defined independently for the parental and offspring individuals (termed 'parental pattern' or 'offspring pattern,' respectively). Both the 'parental pattern' and 'offspring pattern' have six levels (see Tables 2 and 3). However, for downstream analysis, we only analyzed the 'parental patterns' that could generate informative segregation patterns in the offspring: specifically, "Het x Hom-LC", "Het x Hom-HC", and "Het x Het" (3 levels in total). The other categorical variable analyzed represents the haplotype's marker segregation status, and includes three levels: Mendelian error, SD, or normal, as described above. To understand the nature of the relationship between these two categorical variables we visualize their joint occurrences using a mosaic plot. A mosaic plot (Hartigan and Kleiner, 1984) is a visualization tool for contingency tables (i.e., tables of counts). The utility of a mosaic plot is that each cell in a contingency table is represented as a box whose area is proportional to the number of observations, creating an effective graphical display of variable relationships. We represent our data by plotting the 'genotype-coverage' patterns on the y-axis and the haplotype marker segregation status on the x-axis. The height of each 'genotype-coverage' pattern will be proportional to the number of observations for that pattern with respect to all analyzed patterns. The height of the boxes is the same for each 'genotype-coverage' pattern (i.e., row) and is equal to the total counts for that category. Each row can be treated as a histogram with stacked bins. The width of the box is the proportion of observations in a specific row that fall into that category, which has the same column-wise order across rows.

5.3.2 Parental patterns vs. haplotype status

A mosaic plot of the relationship between the 'parental pattern' and the marker segregation status is shown in Figure 3. We found a statistically significant dependence of the two variables (chi^2 test of independence: p-value = 2.2E-16). The low p-value, which is the lowest output given in the R software, is caused by the fact that certain 'parental patterns' do not have observations for all three possbile haplotype status categories. For instance, the 'Het x Het' category is never found to be associated with a marker showing a Mendelian error. The mosaic plots of the two categorical variables were strikingly similar across the three families. Among the three patterns analyzed, the one that proportionally showed that highest number of Mendelian errors was the 'parental pattern' where one

parent is Het and the other a Hom-LC, with a proportion of markers showing Mendelian errors ranging from 142 in family 19 to 200 in family 30.

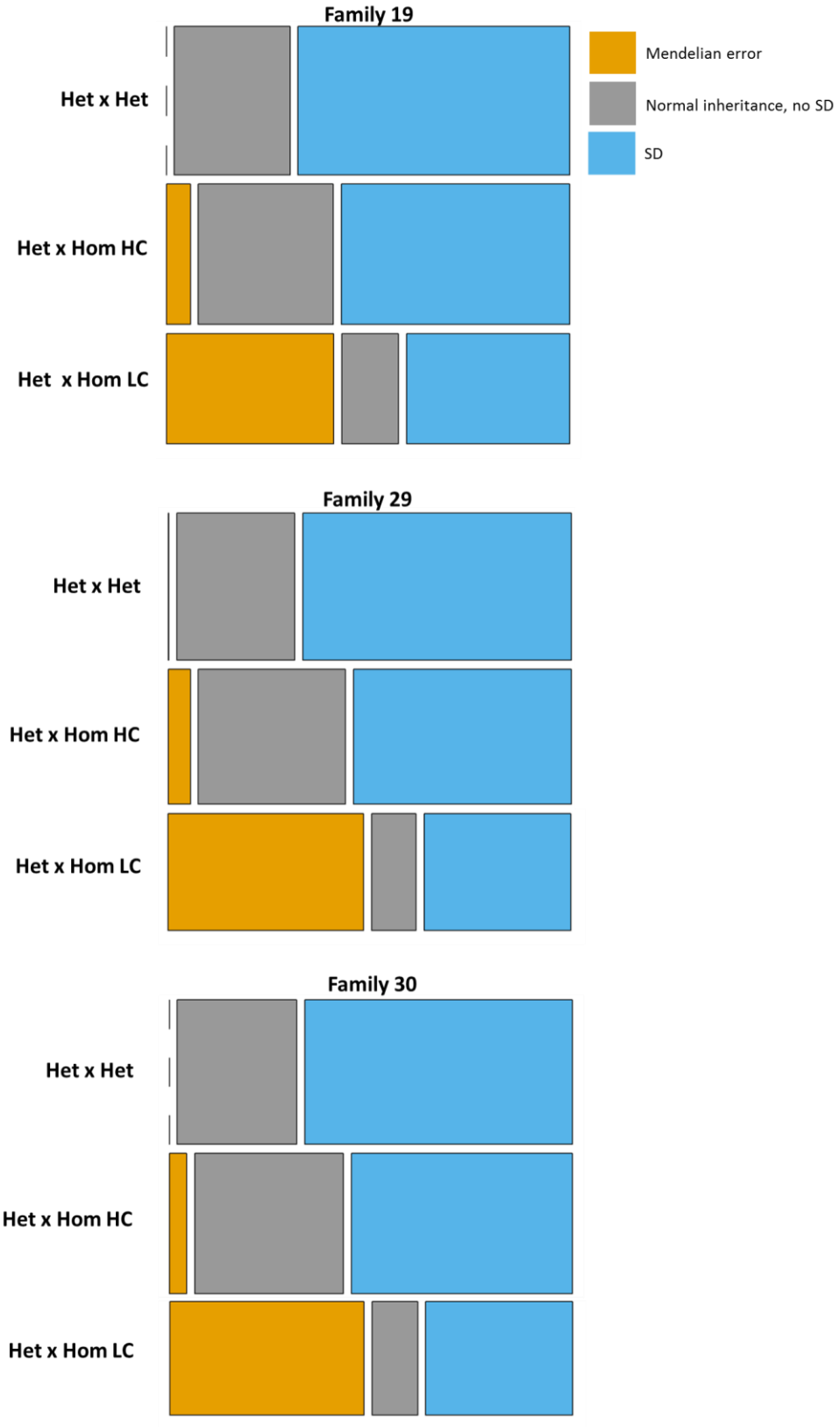


Figure 3. Relationship between the 'parental pattern' (y-axes) and the marker segregation status (x-axes) across three oyster families.

To evaluate whether the high number of Mendelian errors detected in the 'Het x Hom-LC' cross was associated with the segregation of null alleles, we estimated the proportion of offspring categorized as Hom-HC or Hom-LC for the crosses flagged as having Mendelian errors. The rationale is that if the 'Het x Hom-LC' cross carried a null allele, then it should be transmitted to the next generation following the Mendelian rules of inheritance. That is, the putative "AB x AØ" parental cross (represented by the 'Het x Hom-LC' cross) should generate the following genotype proportions in the offspring: 25% AB (i.e., Het), 25% AA (i.e., Hom-HC), 25% AØ (i.e., Hom-LC) and 25% BØ (i.e., Hom-LC). If the genotypes are grouped by 'genotype-coverage' category, then the expectation is 25% Het, 25% Hom-HC and 50% Hom-LC. In Figure 4, it can be observed that the distribution of the proportion of each 'genotype-category' is different, with a slight tendency towards the 'genotype-coverage' expectations (i.e., 50% offspring genotypes = Hom-LC; 25% offspring genotypes = Hom-HC). For the Hom-LC distribution, the proportion of offspring categorized as Hom-LC is skewed towards the expected 50%. On the other hand, the histogram distribution of the Hom-HC is skewed towards a lower proportion of offspring per haplotype marker being Hom-HC.

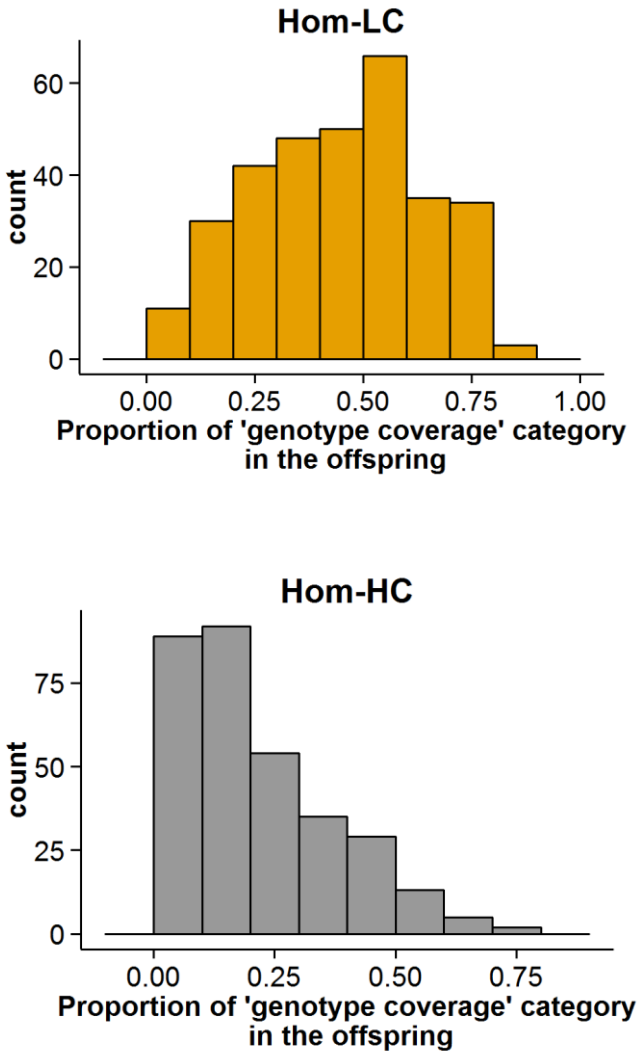


Figure 4. For haplotype markers flagged as showing Mendelian errors, a different distribution of the proportion of homozygote 'genotype categories' (top: Hom-LC; bottom: Hom-HC) is observed in the offspring. In the x-axes the proportion of each 'genotype-coverage' category and in the y-axes the frequency.

Offspring patterns vs. haplotype status

A mosaic plot of the relationship between the 'offspring pattern' and the haplotype marker segregation status is shown in Figure 5. These two variables also showed a significant association (chi^2 test of independence: p-value=2.2E-16). The 'offspring pattern' that showed the highest number of Mendelian errors was the pattern in which the offspring displayed two different types of 'genotype-coverage' categories: Het and Hom-LC. The proportion of Mendelian errors in this category was 38%, 55% 60% in families 19, 29, and 30, respectively. The following category that showed the highest proportion of Mendelian errors per 'offspring pattern' was the one in which all three 'genotype-coverage' categories are found in the offspring: Het, Hom-HC, and Hom-LC. The 'offspring pattern' that showed fewer Mendelian errors was the one where the offspring only displayed the Het and Hom-HC categories. Overall, among the 'offspring patterns', the presence of the Hom-LC (i.e., putative null allele) category in the offspring was associated with a higher number of Mendelian errors.

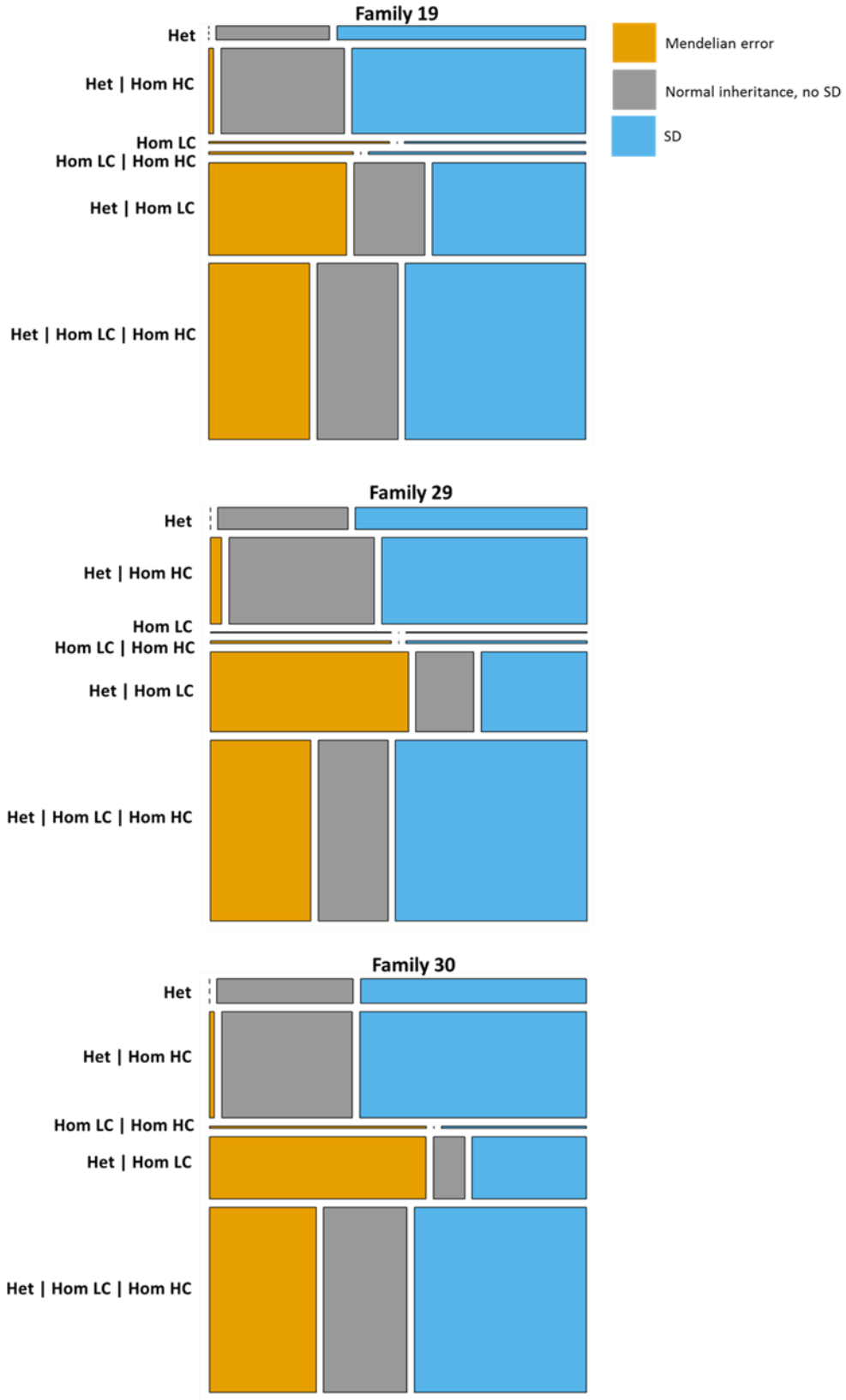


Figure 5. Relationship between the 'offspring pattern' (y-axes) and the marker segregation status (x-axes) across three oyster families.

5.3.4 Parental + Offspring pattern vs. haplotype status

Finally, the 'parental pattern' and 'offspring pattern' were combined into a single categorical variable to explore whether a specific combination was associated with a greater number of Mendelian errors (Figure 6). The highest proportion of Mendelian errors was associated with two combinations of a 'parental pattern' + 'offspring pattern' that fit with the segregation of a putative null allele in the family (patterns marked with a red asterisk in Figure 6). Under the hypothesis that these patterns are reflecting the presence of null alleles, certain conditions – reflected in the number of parental alleles and genotype categories present in the offspring – should be met. A Punnett square explaining these conditions is depicted in Figure 7.

If the 'Het x Hom-LC' (parental pattern) + 'Het | Hom-LC | Hom-HC' (offspring pattern) combination were associated with null alleles, then it would be required, as a first supporting evidence, that (i) both parents share an allele and (ii) three genotype categories are present in the offspring; Hom-LC and Hom-HC should share the same genotype, although the coverage classification should be different (Figure 7A). Of the 163 haplotype markers that showed the abovementioned pattern across families, all fitted the expected offspring genotypes expected to be found in the presence of a segregating null allele when both parents share an allele. Nevertheless, to be taken as evidence of true null alleles, all markers exhibiting the (parental + offspring) patterns that meet these requirements should have been flagged as exhibiting Mendelian errors (100% markers showing Mendelian errors). As observed in Figure 6, however, markers with this pattern, although showing a significant proportion of Mendelian errors, also behave "normally" in some cases, without showing unexpected genotype categories in the offspring. Among the patterns that meet a pattern consistent with a null allele, only 130 (44%), 194 (51%), and 200 (51%) produced a Mendelian error for families 19, 29 and 30, respectively.

On the other hand, if the 'Het x Hom-LC' (parental pattern) + 'Het | Hom-LC' (offspring pattern) was associated with Mendelian errors because of null alleles, then (i) the parents should have no alleles in common and (ii) four different genotype categories should be present in the oyster progeny (Figure 7B). Of the 166 haplotype markers that showed this pattern, only 32 (19%) fitted the number of parental alleles and offspring genotypes expected to be found in the presence of a segregating null allele. The reason for this low fitting to expectation is caused by the fact that we observe less than the four expected

'genotype-coverage' categories in the offspring. A point to be taken into consideration when applying the approach of combining 'parental' and 'offspring' patterns and checking for expected genotype categories in the offspring is that it does not consider the fact that certain genotype categories may be absent because of natural selection (i.e., genotype-specific mortality of larvae).

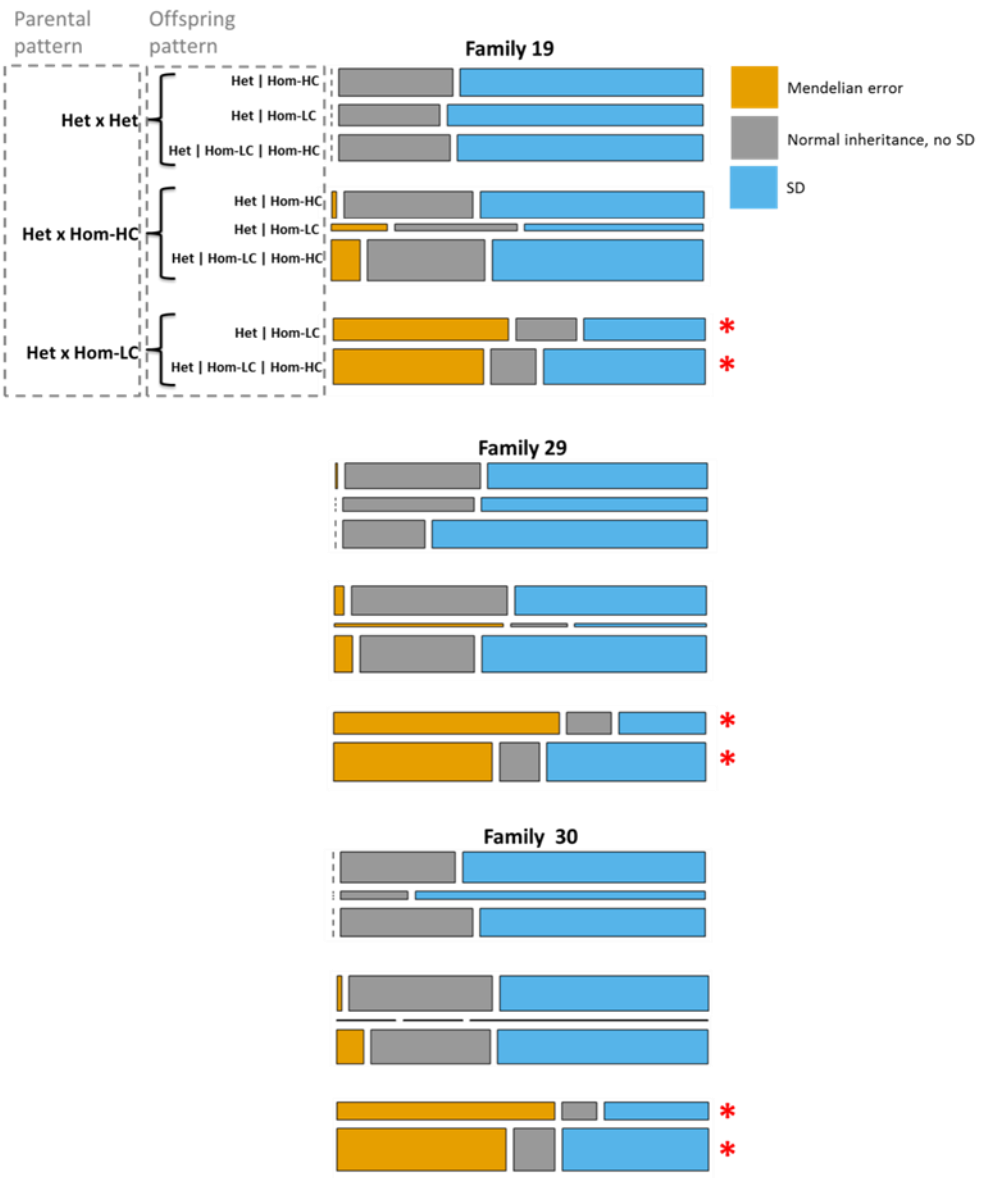


Figure 6. Proportionally higher number of Mendelian errors associated with 'parental pattern' +'offspring pattern' that fit with the segregation of a null allele (patterns with a red asterisk). In the y-axes the different combinations of 'parental patterns' + 'offspring patterns' observed for each family. Only patterns that represent more than 10% of the data are graphed. In the x-axes the marker segregation status that is color coded according to the legend in the top right.

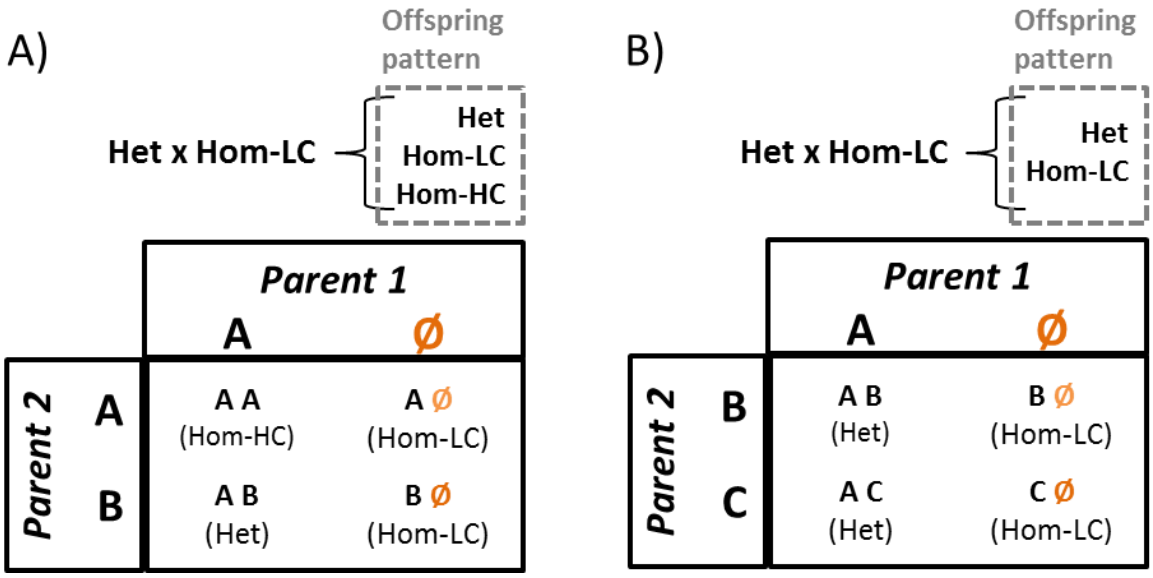


Figure 7. Combinations of 'paternal pattern' + 'offspring pattern' that fit with the segregation of a single null allele (∅) inherited from Parent 1.

5.3.5 Maternal transmission of null alleles

Another approach to evaluate whether Mendelian errors in the RAD-Seq data are caused by null alleles is to examine the stability of the genetic transmission of the Hom-LC loci (marker with a putative null allele). Pacific oyster families 29 and 30 share the same dam. Therefore, if the Hom-LC loci from this dam are in fact null alleles, we would expect them to generate Mendelian errors at the same genomic positions in both families. We cross-referenced all haplotype segregation information for the dam genotypes that belonged to the Hom-LC 'genotype-coverage' category. Of the 665 haplotype markers that were Hom-LC in the abovementioned dam, only 49 generated the same informative crosses for testing this hypothesis [(sire x dam): 'Het x Hom-LC'] at the same loci in both families. The expectation is that at these genomic positions (i.e., loci) Mendelian errors are going to be generated in both families, given that they share the same dam carrying a putative null allele. Among these 49 markers, only 19 (39% → 19/49) displayed Mendelian errors in both families. As an approach to test whether this value was low because of chance, a pairwise comparison of families that shared or did not share the same dam was made. As shown in Table 4, families that share the same dam (families 29 and 30) share more Mendelian errors and

normally behaved markers (i.e., those without Mendelian errors) compared to other possible family pairwise comparisons. Nevertheless, this association is not significant after adjustment for multiple testing (chi^2 test-of-independence: p-value=0.018, p-value threshold=0.016).

Table 4. Number of shared and unique loci showing (i) Mendelian errors or (ii) normal inheritance between two oyster families

Comparison (Family A vs. Family B)	Code	Present only Family A	Present only Family B	Present both
Family 19 vs 29	M. Error	224	310	89
Family 19 vs 30	M. Error	230	304	83
Family 29 vs 30	M. Error	256	244	143
Family 19 vs 29	Normal	502	523	1,139
Family 19 vs 30	Normal	433	581	1,208
Family 29 vs 30	Normal	364	491	1,298

5.4 Discussion

The aim of this Chapter was to test the hypothesis that the large numbers of Mendelian errors detected in the Pacific oyster RAD-Seq data were due to widespread null alleles. Null alleles are a reasonable explanation for the extensive genotype incompatibilities detected in the Pacific oyster RAD-Seq data due to the high sequence and structural polymorphism of the bivalve genome (Zhang et al., 2012b, Sauvage et al., 2007, Rico et al., 2017). The presence of null alleles was examined in our haplotype dataset by incorporating information on locus coverage to each genotyped haplotype marker. Coverage per locus was integrated as a relevant factor into our analysis because we have observed a bimodal coverage per-locus distribution across all individuals and bivalve species analyzed in this thesis (Chapter 3; Figure 6). This bimodal distribution, therefore, appears to be an intrinsic

property of the bivalve genome. One potential explanation for this unusual distribution is that the peak that shows near twice the coverage is the true 'diploid-state' (far right of the distribution), and the peak at half-coverage distance represent genomic markers carrying a null allele. Therefore, this coverage per-locus bimodal distribution may be reflecting a high prevalence of null alleles in the oyster genome, which may, in turn, be generating a large number of null alleles and the consequent high number of Mendelian errors detected in the RAD-Seq data. To test this hypothesis, for each haplotype marker, 'genotype-coverage' patterns were derived from the parental and offspring individuals separately. Patterns were then tested for significant associations with levels of Mendelian errors, as a first approach for establishing a causal relationship and estimating levels of null alleles in bivalve RAD-Seq data.

At an individual level, we found that the proportion of heterozygous genotypes belonging to the low coverage spectrum of the bimodal distribution (Het-LC) is significantly lower compared to the proportion of Hom-LC haplotypes. The non-random association of genotype and position in the bimodal distribution is evidence in favor of the low coverage peak being generated by sequencing only one of the two alleles expected for a diploid species. When individual genotypes were recoded to include locus coverage, and patterns were derived in the parents and offspring, a complex set of pattern was identified. Almost all possible combinations of 'genotype-coverage' categories were observed (Tables 2 and 3). However, certain patterns were substantially more frequent than others. For example, patterns that theoretically would fit with the presence of a null allele – that is, those that contain Hom-LC haplotype categories – exhibit a higher frequency across all Pacific oyster families, representing near 36% and 70% of the 'parental' and 'offspring' patterns, respectively. If these putative null-allele-containing patterns reflect true null alleles, all should produce an incompatibility with Mendelian inheritance (or in other words, a Mendelian error). However, no single null-allele containing pattern showed 100% Mendelian errors in any of the three different types of 'genotype-coverage' patterns evaluated (Figures 3, 4 and 6), revealing that we cannot argue with confidence , under the approach taken, that Mendelian errors are exclusively caused by the segregation of null alleles. Moreover, the fact that families that shared the same dam exhibited a low consistency (~39%) regarding Mendelian errors at putative null maternal alleles (i.e., Hom-LC) is evidence against the presence of null alleles being transmitted from one generation to the next. An interesting observation in early published allozyme work (Lassen and

Turano, 1978, Zouros et al., 1980), which was used against the argument of null alleles as a cause of SD in bivalves, was the absence of null homozygotes (i.e., ØØ). Indeed, if null alleles are segregating at a high frequency in populations, then loci showing SD should also exhibit an increased number of missing genotypes (i.e., the genotypic state of two null alleles being combined). In our study, no significant difference in the overall proportion of missing genotypes was found when pairwise comparisons of parental + offspring patterns were performed. Nevertheless, we did find a significant association between 'genotype-coverage' patterns and haplotype marker status (either 'Mendelian error,' 'SD' or 'Normal inheritance'). Among all possible patterns (parental, offspring or parent + offspring), those that contain the Hom-LC always exhibit a proportionally higher number of Mendelian errors.

The main limitation for discussing potential explanations for the results presented above emerges from the fact that we are using an approach that is error-prone, specifically, because it depends on the interpretation of the origin of the bimodal distribution of coverage. The logic of classifying genotypes by coverage and expecting certain patterns to signify certain biological properties – for example, the Hom-LC is interpreted as a putative null allele – is based on the hypothesis that the lower peak of the distribution corresponds to haplotype markers that carry an allele that is null. Although it is conceivable that bivalves suffer higher rates of null alleles, their magnitude would have to be extreme to generate a distinctive, low coverage peak (Chapter 3; Figure 6). Whether the interpretation of the bimodal coverage distribution is correct would require further testing. For instance, the confirmation of null alleles with a different technology would be a robust approach. Methods are currently available to detect null alleles from SNP array data, which are discovered by an unusual clustering pattern in a Cartesian plot of allele intensities (Carlson et al., 2006, Crooks et al., 2013, Winchester et al., 2008). Unfortunately, our SNP array and RAD-Seq data did not overlap (Chapter 4), limiting the ability to cross-validate putative null alleles.

On the other hand, under the scenario that our approach of integrating locus coverage to genotype information was adequate, and the haplotypes showing lower coverage indeed do tend to represent null alleles, we could suggest that one explanation for not finding a clear causal relationship between putative null alleles and Mendelian errors was the high variability in DNA quality of the oyster samples. To derive 'genotype-coverage' patterns a

threshold of 37x was chosen to separate loci showing high or low coverage (HC and LC, respectively). This value lies approximately in the middle between the two peaks of the bimodal distribution of coverage per locus histograms. A significant issue arises when DNA samples are of poor quality, as they will lead to a lower number of quality reads, which will impact the normalization step. Given that we normalize all individuals by the sample with the fewer number of reads, a lower read output per sample will downsize the sequencing data considerably. Therefore, the major consequence of low-quality samples is that the peaks of the bimodal distribution will become closer and overlap, increasing the risk of misclassification of haplotypes by coverage. The misclassification of individuals may lead to the erroneous assignment of patterns, and spurious associations with haplotype marker status. This situation may be improved by sequencing better quality samples at a higher depth. Another approach would be omitting the read-count normalization step, and instead of using the same cut-off value of coverage for all samples, we could define a specific threshold for each sample in concordance with the sample's particular coverage distribution per locus. By using this approach, bad quality samples would not cause a detrimental effect in the overall analysis, specifically by reducing the threshold criteria, although they would have to be removed from the analysis nonetheless.

In summary, a strong association between patterns containing Hom-LC and high levels of Mendelian errors was found, irrespective if the 'genotype-coverage' category appears in the 'parental-pattern', 'offspring-pattern' or a combination of both. The most likely explanation for this association is null alleles – whether they are inherited or caused by spontaneous (sequence or structural) mutations. The fact that 'genotype-coverage' patterns that would fit with the segregation of a null allele (i.e., patterns containing the Hom-LC category) show no exclusive association with Mendelian errors may be due to the misclassification of genotypes by coverage. This limitation was specifically associated with our experiment and does not reflect a general deficiency of the approach of using coverage to detect putative null alleles. The results presented here suggest that homozgyotes showing low coverage are unreliable for genetic analysis of Pacific oysters, and possibly other bivalve species.

5.5 References

ANDREWS, K. R., GOOD, J. M., MILLER, M. R., LUIKART, G. & HOHENLOHE, P. A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics,* 17**,** 81-92.

ARNOLD, B., CORBETT-DETIG, R. B., HARTL, D. & BOMBLIES, K. 2013. RAD-seq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology,* 22**,** 3179-3190.

BROQUET, T. & PETIT, E. 2004. Quantifying genotyping errors in noninvasive population genetics. *Molecular Ecology,* 13**,** 3601-3608.

CARLSON, C. S., SMITH, J. D., STANAWAY, I. B., RIEDER, M. J. & NICKERSON, D. A. 2006. Direct detection of null alleles in SNP genotyping data. *Human Molecular Genetics,* 15**,** 1931-1937.

CATCHEN, J., HOHENLOHE, P. A., BASSHAM, S., AMORES, A. & CRESKO, W. A. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology,* 22**,** 3124-3140.

CATCHEN, J. M., AMORES, A., HOHENLOHE, P., CRESKO, W. & POSTLETHWAIT, J. H. 2011. Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes|Genomes|Genetics,* 1**,** 171-182.

CHAPUIS, M.-P. & ESTOUP, A. 2007. Microsatellite Null Alleles and Estimation of Population Differentiation. *Molecular Biology and Evolution,* 24**,** 621-631.

CONRAD, D. F., ANDREWS, T. D., CARTER, N. P., HURLES, M. E. & PRITCHARD, J. K. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetic,* 38**,** 75-81.

COOKE, T. F., YEE, M., MUZZIO, M., SOCKELL, A., BELL, R. & CORNEJO, O. E. 2016. GBStools: a statistical method for estimating allelic dropout in reduced representation sequencing data. *PLoS Genetics,* 12, e1005631.

CROOKS, L., CARLBORG, Ö., MARKLUND, S. & JOHANSSON, A. M. 2013. Identification of Null Alleles and Deletions from SNP Genotypes for an Intercross Between Domestic and Wild Chickens. *G3: Genes|Genomes|Genetics,* 3**,** 1253-1260.

DAKIN, E. E. & AVISE, J. C. 2004. Microsatellite null alleles in parentage analysis. *Heredity,* 93**,** 504-509.

DAVEY, J. W., CEZARD, T., FUENTES-UTRILLA, P., ELAND, C., GHARBI, K. & BLAXTER, M. L. 2013. Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology,* 22**,** 3151-3164.

DAVEY, J. W., HOHENLOHE, P. A., ETTER, P. D., BOONE, J. Q., CATCHEN, J. M. & BLAXTER, M. L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics,* 12, 499-510.

GAUTIER, M., GHARBI, K., CEZARD, T., FOUCAUD, J., KERDELHUÉ, C. & PUDLO, P. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Moleular Ecology,* 22, 3165-3178 .

HARE, M. P., KARL, S. A. & AVISE, J. C. 1996. Anonymous nuclear DNA markers in the American oyster and their implications for the heterozygote deficiency phenomenon in marine bivalves. *Molecular Biology and Evolution,* 13**,** 334-345.

HARRANG, E., LAPEGUE, S., MORGA, B. & BIERNE, N. 2013. A High Load of Non-neutral Amino-Acid Polymorphisms Explains High Protein Diversity Despite Moderate Effective Population Size in a Marine Bivalve With Sweepstakes Reproduction. *G3: Genes|Genomes|Genetics,* 3**,** 333-341.

HARTIGAN, J. A. & KLEINER, B. 1984. A Mosaic of Television Ratings. *The American Statistician,* 38**,** 32-35.

HEDGECOCK, D., LI, G., HUBERT, S., BUCKLIN, K. & RIBES, V. 2004. Widespread null alleles and poor cross-species amplification of microsatellite DNA loci cloned from the Pacific oyster, Crassostrea gigas. *Journal of shellfish research,* 23**,** 379-385.

ILUT, D. C., NYDAM, M. L. & HARE, M. P. 2014. Defining Loci in Restriction-Based Reduced Representation Genomic Data from Nonmodel Species: Sources of Bias and Diagnostics for Optimal Clustering. *BioMed Research International,* 2014**,** 9.

KOHLER, JARED R. & CUTLER, DAVID J. 2007. Simultaneous Discovery and Testing of Deletions for Disease Association in SNP Genotyping Studies. *American Journal of Human Genetics,* 81**,** 684-699.

LASSEN, H. H. & TURANO, F. J. 1978. Clinal variation and heterozygote deficit at the lap-locus in Mytilus edulis. *Marine Biology,* 49**,** 245-254.

LAUNEY, S. & HEDGECOCK, D. 2001. High genetic load in the Pacific oyster Crassostrea gigas. *Genetics,* 159**,** 255-265.

LEMER, S., ROCHEL, E. & PLANES, S. 2011. Correction Method for Null Alleles in Species with Variable Microsatellite Flanking Regions, A Case Study of the Black-Lipped Pearl Oyster Pinctada margaritifera. *Journal of Heredity,* 102**,** 243-246.

MACAVOY, E. S., WOOD, A. R. & GARDNER, J. P. A. 2008. Development and evaluation of microsatellite markers for identification of individual Greenshell™ mussels (Perna canaliculus) in a selective breeding programme. *Aquaculture,* 274**,** 41-48.

MCCARROLL, S. A., KURUVILLA, F. G., KORN, J. M., CAWLEY, S., NEMESH, J., WYSOKER, A., SHAPERO, M. H., DE BAKKER, P. I. W., MALLER, J. B., KIRBY, A., ELLIOTT, A. L., PARKIN, M., HUBBELL, E., WEBSTER, T., MEI, R., VEITCH, J., COLLINS, P. J., HANDSAKER, R., LINCOLN, S., NIZZARI, M., BLUME, J., JONES, K. W., RAVA, R., DALY, M. J., GABRIEL, S. B. & ALTSHULER, D. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics,* 40**,** 1166-1174.

MCGOLDRICK, D. J., HEDGECOCK, D., ENGLISH, L. J., BAOPRASERTKUL, P. & WARD, R. D. 2000. The transmission of microsatellite alleles in Australian and North American stocks of the Pacific oyster (Crassostrea gigas): Selection and null alleles. *Journal of Shellfish Research,* 19**,** 779-788.

MORVEZEN, R., CORNETTE, F., CHARRIER, G., GUINAND, B., LAPÈGUE, S., BOUDRY, P. & LAROCHE, J. 2013. Multiplex PCR sets of novel microsatellite loci for the great scallop Pecten maximus and their application in parentage assignment. *Aquatic Living Resources,* 26**,** 207-213.

PALLAVICINI, A., DEL MAR COSTA, M., GESTAL, C., DREOS, R., FIGUERAS, A., VENIER, P. & NOVOA, B. 2008. High sequence variability of myticin transcripts in hemocytes of immune-stimulated mussels suggests ancient host–pathogen interactions. *Developmental & Comparative Immunology,* 32**,** 213-226.

PINO-QUERIDO, A., ÁLVAREZ-CASTRO, J. M., VERA, M., PARDO, B. G., FUENTES, J. & MARTÍNEZ, P. 2015. A molecular tool for parentage analysis in the Mediterranean mussel (Mytilus galloprovincialis). *Aquaculture Research,* 46**,** 1721-1735.

R CORE TEAM 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

REECE, K. S., RIBEIRO, W. L., GAFFNEY, P. M., CARNEGIE, R. B. & ALLEN, J. S. K. 2004. Microsatellite Marker Development and Analysis in the Eastern Oyster (Crassostrea virginica): Confirmation of Null Alleles and Non-Mendelian Segregation Ratios. *Journal of Heredity,* 95**,** 346-352.

REN, J., LIU, X., ZHANG, G., LIU, B. & GUO, X. 2009. "Tandem duplication-random loss" is not a real feature of oyster mitochondrial genomes. *BMC Genomics,* 10**,** 84.

RICO, C., CUESTA, J. A., DRAKE, P., MACPHERSON, E., BERNATCHEZ, L. & MARIE, A. D. 2017. Null alleles are ubiquitous at microsatellite loci in the Wedge Clam (Donax trunculus). *PeerJ,* 5**,** e3188.

SAUVAGE, C., BIERNE, N., LAPÈGUE, S. & BOUDRY, P. 2007. Single Nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster Crassostrea gigas. *Gene,* 406**,** 13-22.

SEROUSSI, E., GLICK, G., SHIRAK, A., YAKOBSON, E., WELLER, J. I., EZRA, E. & ZERON, Y. 2010. Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs. *BMC Genomics,* 11**,** 673.

WINCHESTER, L., NEWBURY, D. F., MONACO, A. & RAGOUSSIS, J. 2008. Detection, breakpoint identification and detailed characterisation of a CNV at the FRA16D site using SNP assays. 123**,** 322-332.

YU, Z. N., WEI, Z. P., KONG, X. Y. & SHI, W. 2008. Complete mitochondrial DNA sequence of oyster Crassostrea hongkongensis -a case of "Tandem duplication-random loss" for genome rearrangement in Crassostrea? *BMC Genomics,* 9, 477.

ZHAN, A.-B., BAO, Z.-M., HUI, M., WANG, M.-L., ZHAO, H.-B., LU, W., HU, X.-L. & HU, J.-J. 2007. Inheritance pattern of EST-SSRs in self-fertilized larvae of the bay scallop Argopecten irradians. *Annales Zoologici Fennici,* 44**,** 259-268.

ZHANG, G., FANG, X., GUO, X., LI, L., LUO, R., XU, F., YANG, P., ZHANG, L., WANG, X., QI, H., XIONG, Z., QUE, H., XIE, Y., HOLLAND, P. W. H., PAPS, J., ZHU, Y., WU, F., CHEN, Y., WANG, J., PENG, C., MENG, J., YANG, L., LIU, J., WEN, B., ZHANG, N., HUANG, Z., ZHU, Q., FENG, Y., MOUNT, A., HEDGECOCK, D., XU, Z., LIU, Y., DOMAZET-LOSO, T., DU, Y., SUN, X., ZHANG, S., LIU, B., CHENG, P., JIANG, X., LI, J., FAN, D., WANG, W., FU, W., WANG, T., WANG, B., ZHANG, J., PENG, Z., LI, Y., LI, N., WANG, J., CHEN, M., HE, Y., TAN, F., SONG, X., ZHENG, Q., HUANG, R., YANG, H., DU, X., CHEN, L., YANG, M., GAFFNEY, P. M., WANG, S., LUO, L., SHE, Z., MING, Y., HUANG, W., ZHANG, S., HUANG, B., ZHANG, Y., QU, T., NI, P., MIAO, G., WANG, J., WANG, Q., STEINBERG, C. E. W., WANG, H., LI, N., QIAN, L., ZHANG, G., LI, Y., YANG, H., LIU, X., WANG, J., YIN, Y. & WANG, J. 2012b. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature,* 490**,** 49-54.

ZOUROS, E., SINGH, S. M. & MILES, H. E. 1980. Growth Rate in Oysters: An Overdominant Phenotype and Its Possible Explanations. *Evolution,* 34**,** 856-867.

CHAPTER 6

Characterisation of *de novo* mutations in Greenshell™ mussel (*Perna canaliculus*) full-sib families

6.1 Introduction

One of the most important factors that predict genetic diversity is the life-history of a species (Ellegren and Galtier, 2016). In a comparative genome-wide survey across 76 non-model animal species, the estimates of synonymous nucleotide diversity (π_s) were similar within taxonomic families, supporting the theory that a species' biology is a good predictor of diversity (~73% of variance in π_S explained) (Romiguier et al., 2014). The primary life-history traits explaining variations in polymorphism between species are propagule size and fecundity, both a reflection of parental investment in reproduction. As highly fecund, sessile, broadcast spawners, marine bivalves are considered extreme examples of species with low parental investment (termed 'r-strategists'). For sessile organisms, favouring offspring quantity over quality appears to be beneficial in the face of changing environmental conditions, where fitness can have a low adaptive value. However, the question of whether the high number of offspring generated by r-strategists is created at the expense of individual viability is an aspect that has rarely been considered in animal research.

Normal cells replicate their DNA with remarkable fidelity, as a result of key processes that include nucleotide selectivity, proofreading and DNA mismatch repair (Arana and Kunkel, 2010). If defects are present in any of these genome maintenance systems, then increased mutation rates are expected. For instance, mutations in the mismatch repair mechanism of humans can cause hereditary colon cancer via increased spontaneous mutations (Peltomaki, 2001). Hence, it is plausible that during the process of generating the large amounts of gametes required by r-strategists, a higher frequency of germline mutations is being produced due to the greater demand for the DNA replication machinery. A high germline mutation rate may consequently lead to increased genome instability of the offspring of highly fecund species, which may ultimately affect the overall survival (viability) because of endogenous factors. In bivalves, studies of genome inheritance are starting to reveal that some of the unusual genetic properties observed in this group of species, such as extensive marker segregation distortion, are triggered by endogenous factors that underlie differential larval survival (Bierne et al., 1998, Plough and Hedgecock, 2011, Plough et al., 2016, Launey and Hedgecock, 2001). Consequently, fundamental 'unsolved' aspects of the genetics and biology of marine bivalves can be better understood by analysing the characteristic stage of species with indirect development – the larva.

As discussed in Chapter 1 a popular suggested reason for the genome-wide patterns of SD of marine bivalves is viability selection acting on deleterious mutations during the larval stages of the bivalve life-cycle (Bierne et al., 1998, Launey and Hedgecock, 2001, Plough and Hedgecock, 2011, Plough et al., 2016). Indeed, bivalve larvae experience substantial mortality during early-life history – with field estimates reaching up to 99% (Thorson, 1950). While these high mortalities were traditionally thought to be caused primarily by exogenous factors such as predation, food availability or physical factors (Thorson, 1966, Gosselin and Qian, 1997, Rumrill, 1990), there is growing evidence that endogenous factors such as lethal mutations, a suggested intrinsic property of marine bivalves (see, for a review, Plough (2016)), may also play a major role. The evidence that SD in bivalves is caused by early selection has been provided by studies that genotyped single larva of experimental pair crosses. Bierne et al. (1998) sampled day 1 (early larval stage), day 10 (late larval stage), and day 70 (juvenile, sessile stage) flat oyster offspring, and showed that 1-day-old larvae typed with 4 microsatellite markers conformed to Mendelian segregation ratios. The offspring from the last temporal sample (day 70) exhibited variable levels of SD, which varied amongst markers and families. This temporal analysis allowed the detection of the time at which putative strong selection was occurring – mainly at metamorphosis – along with evidence for its major role in determining the SD patterns that are observed later in life. A study by Launey and Hedgecock (2001) supported these findings, and by genotyping larval offspring at a higher marker density (n total = 19 microsatellites), mapped the deleterious mutations using distorted proxy markers in linkage disequilibrium. Using this approach, they estimated that the wild founders of inbred Pacific oyster lines carried a minimum of 8-14 highly deleterious recessive mutations. The segregation of these lethal variants in the offspring was estimated to be responsible for 80% of the larval mortalities. Subsequent studies employed a first generation Pacific oyster linkage map (~80 microsatellites) to map viability QTL (Plough and Hedgecock, 2011, Plough, 2012), suggesting similar numbers of segregating deleterious mutations, which mainly acted during the larval stage and near metamorphosis. The substantial early-life history mortalities and the extensive SDs of bivalves, which have been demonstrated to arise during larval stages, have been interpreted as evidence for a high genetic load in marine bivalves (Plough, 2016). Although a (deleterious) mutational load has been hypothesised as the main contributing factor to this genetic load, the source of these mutations and direct evidence for a high mutation rate in bivalves has not yet been reported.

In addition to the fundamental evolutionary questions that may be associated with a potentially high level of mutation, they must be accounted for in selective breeding programs for these species. De novo mutation rates in other farmed animals are so low that they generally do not contribute significantly to genetic variation underpinning complex traits of economic importance. However, it is plausible that in addition to the putative viability loci segregating in bivalve families, the high mutation rates inferred for bivalve species may also contribute to genetic variation in all heritable traits. Therefore, quantification and characterization of these phenomena are important for informing strategies for selective breeding, and optimal use of genomic tools to enhance aquaculture production of these critically important food production species.

The Greenshell™ mussel (*Perna canaliculus*) is New Zealand's main aquaculture species (FAO, 2005). With most of the spat (juveniles) being traditionally collected from the wild, the establishment of a commercial hatchery in 2015 marked the beginning of selective breeding for trait improvement in this endemic species. To support Greenshell™ mussel breeding programs, different technologies have been developed, including oocyte cryopreservation methods (Adams et al., 2009), parentage assignment tools (MacAvoy et al., 2008), and high-density culture systems (Ragg et al., 2010). Additional technologies that would allow for the genetic improvement of key traits in these mussel stocks would be high-density genetic markers, which could be applied for parental assignment, the creation of high-density linkage maps and genome-enabled selection. Nevertheless, the utility of DNA marker technology will depend on our understanding of the stability of genetic transmission in mussels, particularly in view of the high quantity of apparent Mendelian errors observed previously in this thesis (Chapter 5). Moreover, this situation can be further complicated by the proposed argument that bivalves exhibit high mutation rates, which may lead to Mendelian errors, segregation distortion and novel contribution to trait variation that is not typically accounted for in quantitative genetics models underpinning selective breeding models.

The aim of this Chapter was to identify, quantify and characterize *de novo* mutations in five full-sibling families of the Greenshell™ mussel. The mussel families were created in replicate and reared separately until settlement. Samples from the parents, their gametes, and pools of their offspring at different larval stages were sequenced using RAD-Seq. To examine if mutations originated during gamete formation, we compared the genotypes of

the parents with their respective gametes. To determine if mutations appeared post-zygotically, we searched for novel alleles in pools of larval offspring and estimated at which larval stage the novel alleles appeared in the temporal samples of larvae. Additionally, as an indirect indication of a possible association of putative *de novo* mutations with SDs, changes in allele frequencies during larval development were monitored. Using these families and genetic marker resources, a characterization of the timing and origin of *de novo* mutation was provided.

6.2 Material and Methods

6.2.1 Pair-crosses

Five Greenshell™ (*Perna canaliculus*) mussel families consisting of two full-siblings and three half-sibling families were produced in 2016 (see Chapter 2). The dams used in this study were derived from a line selected for growth, whereas sires were randomly sampled from a wild population from Nelson, New Zealand. Families were created in replicate, with gametes from the same parents being fertilized twice, and each reared separately. Pedigree information and family labels are shown in Figure 1.

Given that the primary aim of the study was to detect *de novo* mutations in pools of larval offspring, the experimental layout and procedures were designed to minimize chances of cross contamination of individuals in the hatchery. For instance, the physical distance of the experimental tanks was maximized at the different stages of the larval rearing procedure, to avoid the remote occasion of cross-contamination of larvae from different families. Additionally, as part of the larval rearing procedure, the Cawthron Ultra-Density Larval rearing (CUDL) system (Ragg et al., 2010) has to be cleaned every other day. Distinct cleaning materials were used for each set of family replicates. The whole larval rearing process, from fertilisation to settlement was carried out at the Cawthron Aquaculture Park, Nelson, New Zealand, where families were grown in a closed system with UV-treated and filtered sea water, addressing the possibility that foreign larvae may enter the experiment through the water supply (see Chapter 2 for details).

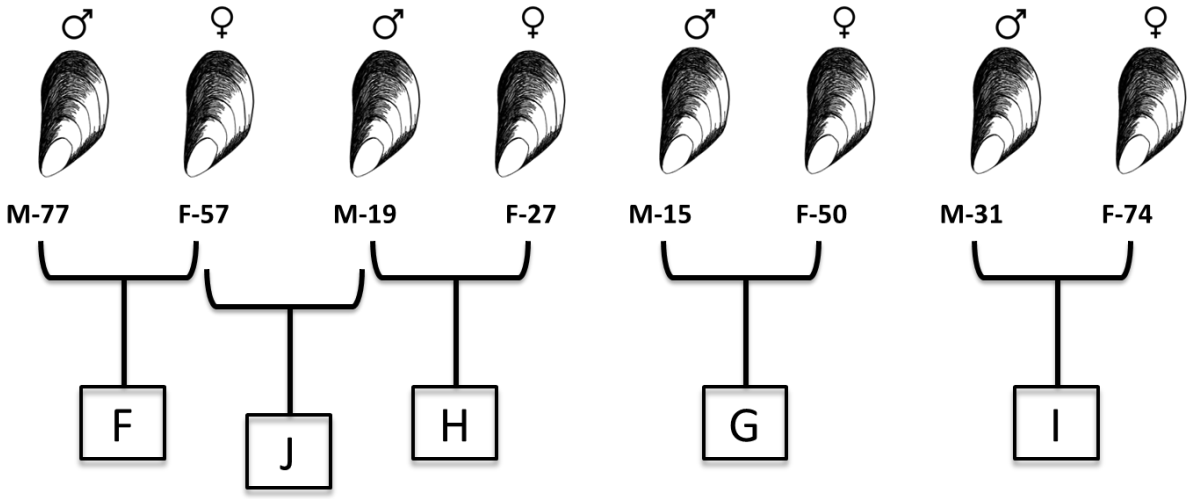


Figure 1. Description of the pedigrees used in this study. The ID of the male and female individuals used as parents in this study is indicated below each mussel picture. M stands for 'male' and F stands for 'female'. Mussel drawing is taken from http://www.kentandessex-ifca.gov.uk

6.2.2 Collection of samples

The aims of this Chapter were to (i) detect, quantify and characterize *de novo* mutations in Greenshell™ mussels by sequencing genomic DNA from nuclear family samples; if present the subsequent aim was to (ii) identify the sources (i.e., germline or post-zygotic) of these novel alleles, and (iii) monitor significant changes in allele frequencies, as a means of determining the timing of SD in larval development.

Different sets of samples were collected to address the different aims proposed. To detect *de novo* mutations and monitor allele frequencies during development (aims i and iii), tissue from all parents (4 males and 4 females) and pools of their larval offspring at different time points of development: 20 hours (h) (trochophore), 40h (early veliger), 4 days (d) (veliger), 8d (veliger) and 12d (late veliger/pediveliger) were sampled. Regarding the origin of *de novo* mutations (aim ii), to identify whether the source was the germline, samples of sperm and oocytes (henceforth termed 'eggs') were taken from each parent. On

the other hand, to detect mutational events arising post-zygotically, previously sampled pools of larvae and parents (as used for aim iii) were utilized.

6.2.3 DNA extraction

DNA was extracted from pools of gametes and offspring with protocols that depend on the type of biological sample (see Chapter 2 - section 2.1.2). A major challenge in pooled sequencing is to get an accurate allele frequency estimation (Andrews et al., 2016). Moreover, when looking for polymorphisms and *de novo* mutations the frequency at which they occur will determine the number of individuals to sample, and the power of the analysis. In the absence of preliminary information regarding the frequency of putative *de novo* mutations in mussels, a high number of individual larvae were sampled at each time-point, estimated at between 5,000 and 10,000.

6.2.4 RAD-Seq library preparation

RAD libraries were prepared following the protocol described in section 2.1.3 (Chapter 2). In total, the genomic DNA from each of the 8 parental individuals, 8 pools of their respective gametes, and 48 pools of larval offspring were digested separately with a *SbfI* restriction enzyme. To detect any potential technical sources of variation, two technical replicates – consisting of the same DNA sample processed independently during library preparation – were included in the analysis. Each individual sample was barcoded with a unique combination of 11 P1 and 6 P2 adapter sequences. All samples were pooled in equimolar amounts and sequenced in a single lane of an Illumina Hiseq2500 platform (2x125bp) at the Edinburgh Genomics facility, Edinburgh, UK.

6.2.5 De novo assembly

De novo assembly was performed using the pseudo-reference pipeline developed and evaluated in the previous Chapter (section 3.2.5, Chapter 3). Briefly, the assembly of Greenshell™ mussel single-end reads was performed using Stacks (Catchen et al., 2011, Catchen et al., 2013) with the following parameters: reads were assembled into RAD-loci

with the *ustacks* module using a minimum stack depth of 10, and a maximum number of mismatches between stacks (or the two putative alleles of an individual) of 6. In addition, we allowed for the presence of indel variation enabling a '--gapped-alignment'. The total number of individual RAD-loci identified across family members was then merged into a master file and clustered by similarity (using an identity threshold of 94%) into a representative set of family-specific loci using CD-hit (Fu et al., 2012). A preliminary evaluation of the *de novo* assembly indicated that building a reference pseudo-genome from loci assembled across mussel families was not conducive to the analysis. The number of common RAD loci identified across families was significantly lower than expected when compared to an approach where family-specific RAD loci were defined (example in Appendix, Figure 1). The cause of this observation is unknown, although it may be related to the high levels of sequence and structural polymorphism observed in bivalve genomes (Zhang et al., 2012), which may be limiting to some extent the identification of homologous loci across unrelated individuals. Since the primary aim of the study was to detect mutations, the number of genomic regions interrogated within each family was maximised by conducting *de novo* assembly and variant calling within each family separately. Quality-filtered reads from each sample were aligned to their respective family pseudo-reference with BWA under highly relaxed parameter values (Li and Durbin, 2009) (full BWA command: bwa aln -l 210 -o 2 -n 0.01 -e 10 -d 12). These parameters were used to account for the possibility that samples within the same nuclear family may be unusually divergent under the hypothesis of a high mutation rate. The sequence alignments were converted to binary format (BAM) files, sorted and indexed using SAMtools v1.4 (Li et al., 2009). Prior to data analysis, the alignment files were filtered for low-quality bases (BQ<15) and reads with low mapping quality (MQ<20).

At this stage, two aligned datasets were created for downstream analyses: one that contained all parents and their respective gametes (*parent-gamete* dataset) and another that contained all parents plus their offspring sampled at different larval stages (*parent-offspring* dataset). These files were processed via three different methods depending on the aims that were established, as outlined below:

6.2.6 Prediction of an 'expected' mussel progeny

To study SD occurring at early developmental stages, larval offspring samples were taken at different time points from age 20h until 2 weeks after settlement. However, this experimental design is not able to capture segregation distortion if it occurred between fertilization and the first sampling point. As an approach to estimate levels of distortion at this very early stage, the hypothetical genetic pool of the offspring populations was mimicked based on the gamete genotype data according to Mendelian expectations, assuming (i) all parental gametes had an equal chance of fertilization and (ii) no post-zygotic selection occurred. For each family replicate, post-filtered sequencing reads from eggs and sperm were combined into a single file (mimicking alleles that are randomly brought together at fertilization in a marine environment). Given each read in the gamete (pool) sample should correspond exactly to an individual gamete, male and female gamete pools have to be normalized by read number before merging them. This 'virtual pool of embryos' can be considered to represent a time zero sampling point and was used to extrapolate our analysis to a theoretical earlier stage. The 'virtual pool of embryos' were created for each parent pair and then included in the *parent-offspring* datasets.

6.2.7 Processing of *parent-offspring* dataset

Two approaches were utilized for variant calling depending on the question under study:

6.2.7.1 Changes in allele frequencies during larval development

An analysis of temporal changes in allele frequencies was performed by comparing two consecutive larval stages. The "virtual" pooled offspring samples (see above) were treated as the time-point zero in the analysis. Significant differences in allele frequencies between developmental stages were detected using a Fisher exact test implemented in PoPoolation2 (Kofler et al., 2011), by using a sliding-window approach (window size = RAD-loci length = 100bp). (RAD) loci that showed a minimum of 15x (--min-coverage), a maximum of 200x (--max-coverage), and a minimum allele count of 3 (--min-count) were included in the analyses. Significance levels were adjusted for multiple comparisons using a Bonferroni correction (α=0.05). After detecting significantly distorted loci within families, we aimed to

evaluate whether the same loci were responsible for segregation distortion (if present) at each other developmental stage within families.

6.2.7.2 Estimation of post-zygotic mutations in pools of mussel larvae

To monitor single-point post-zygotic mutations arising at different larval development stages, variants were called using the parents and the pooled offspring samples utilizing VarScan2 (Koboldt et al., 2012). VarScan employs a heuristic approach to identify variants, with the main parameters controlling for sensitivity and specificity being the minimum coverage (--min-coverage) and minimum variant frequency (-min-var-freq). Variants were called using the mpileup2snp command, with the following parameter value settings: a minimum of 3 reads were required to call a variant (--min-reads2), the minimum variant allele frequency threshold of the variant was set to 0.01 (--min-var-freq), a minimum coverage of 20x was required to process a site (--min-coverage), the strand bias filter was disabled (--strand-filter=0), and the p-value threshold to call a variant was set to 0.3. The latter p-value threshold was low stringency to enable the software to call a 'heterozygous genotype' with low evidence from the variant allele, which would be the case under the presence of *de novo* mutations occurring at a low frequency in pooled samples of sibling larvae.

6.2.8 Processing of *parent-gamete* datasets

Two *parent-gamete* datasets were created depending on the objective.

6.2.8.1 Estimation of germline mutations

To detect germline *de novo* mutations in male and female mussel gametes, pairs of samples (parent vs. pool of gametes) were compared using a VarScan2 command designed to distinguish amongst (i) germline (inherited) variants, (ii) loss of heterozygosity (LOH) events or (iii) somatic mutations in paired tumor/normal samples. These categories are assigned based on the presence/absence of alleles in a pair of sequenced samples. For a given SNP marker, if the genotype of the 'normal' and so-called 'tumor' sample match, then it is categorized as a germline inherent variant. For instances in which the 'normal' sample is genotyped as a heterozygous and the 'tumor' homozygous, the mutation would be considered to be caused by a chromosomal loss event, and therefore would be classified as

LOH. On the other hand, if a 'normal' sample is homozygous and the 'tumor' is heterozygous, then it would be considered a somatic mutation.

For the *parent-gamete* dataset, the focus was on the commands used to detect somatic mutations, as they resemble the pattern expected for a *de novo* mutation – i.e., alleles present in gametes and absent in the parent sample. The VarScan *somatic* command searches for variants in each sample according to user-defined parameter values. Once a variant is called in either sample, the genotypes and supporting data (read counts for the reference and variant allele) for both samples are compared with a Fisher's Exact test. For each parent-gamete pairwise comparison, the parents were treated as the 'normal' sample and their respective gametes as the 'tumor' sample. De novo mutations were identified in the Greenshell™ mussel parent-gamete pairs using the following VarScan parameters: --min-coverage 15, --min-var-freq 0.05, --somatic-p-value 0.01, --strand-filter 0 and --min-avg-qual 20. No strand bias filter was applied because of the nature of RAD-Seq technique. An additional filter was applied to remove SNPs that were located near indels, as they may be false positives derived from local misalignment. A subset of high-confidence (H-C) variants were obtained by running the *processSomatic* command requiring a variant allele frequency (VAF) equal to 0% in the 'normal' sample (i.e., parents) and higher than 2% VAF in the 'tumor' (i.e., pool of gametes) sample. A preliminary evaluation of the distribution of the number of SNPs across the single end read showed an enrichment of polymorphic variants at three specific nucleotide positions: 7bp, 70bp, and 97bp (Appendix, Section 1: Figure 2). The expectation is a homogeneous distribution of genuine SNP markers along the read. It was hypothesized that the observed pattern might have been caused by a systematic error being introduced during the sequencing process on the Illumina platform (Nakamura et al., 2011). The sequencing facility was consulted for a possible explanation. Due to the fact that the internal control of the sequencing run (PhiX virus) did not exhibit a peak at the aforementioned positions, the reason for the non-random distribution of SNPs remains unknown. To avoid potential biases, markers identified at these positions were removed and not taken into account when estimating numbers of putative *de novo* mutations.

6.2.8.2 Transmission ratio distortion

Transmission ratio distortion (TRD) is observed when one of the alleles from a parent is preferentially transmitted to the offspring. This unequal segregation of alleles can be detected by testing for a significant departure from the Mendelian inheritance ratio of 0.5 per parental allele (Pardo-Manuel de Villena and Sapienza, 2001). To generate this dataset, variants were called from parental samples and pools of gametes utilizing the same pipeline that was used for estimating post-zygotic mutations in larvae (see section 6.2.7.2). From this dataset, the allele ratio for parental heterozygous alleles was calculated and compared to the allelic ratio of their respective gametes. The allele ratios from both pairs of samples (parent vs. gametes) were evaluated for a deviation from a 1:1 ratio using a chi^2 goodness-of-fit-test. The nominal significance threshold was 0.05.

6.2.9 Evaluation of *de novo* mutations: genome measures used for quality control

The transition/transversion ratio (Ti/Tv) was used as a metric to distinguish true mutations from technical errors, and is generally used as a quality control procedure for SNP discovery. The Ti/Tv ratio of neutral mutations or random sequencing errors should be close to ~0.5 (case in which all nucleotides have an equal probability of mutating to another state).

6.3 Results

The number of paired-end reads obtained per sample ranged from 1,411,820 to 8,907,054, with an average of 4,146,642 (SD = 1,611,394). After the quality control filtering and read alignment, it was observed that the sequencing depths per locus vary significantly according to the biological sample (Appendix, Section 1: Figure 3). Comparatively, the egg and trochophore samples had a lower average locus coverage (<100x across all families). The effect of this low coverage on the different analyses is negligible, as at most they lead to the underestimation of the frequency of putative mutations, which herein we observe at a high frequency in both eggs and trochophore. The coverage plots by sample also confirm the results presented in previous Chapters (Chapter 3) for other bivalve species, as a

bimodal distribution of two distinct coverage categories in the Greenshell™ mussel genome was observed.

6.3.1 Temporal analysis

A Fisher's exact test was used to detect significant changes in allele frequencies between pools of larval offspring from five Greenshell™ mussel families. We found a substantial change in allele frequencies throughout larval development. The results of the temporal analysis are expressed as the percentage of loci examined that show a statistically significant difference in allele frequency between two consecutive larval stages. In total, 25% (567/2,264), 29% (378/1,319), 23% (1,067/4,602), 24% (755/3,118) and 22% (951/4,247) of loci (i.e., 100bp regions) became distorted at least once in the larval offspring of families F, G, H, I and J, respectively. Across families, the same tendency was observed, with a higher proportion of loci showing significant changes in allele frequencies arising at earlier stages (Figure 1). Our first sampling time point (time zero in larval development) is represented by a 'virtual' sample created by merging the sequencing files of the parental sperm and eggs from each family unit; mimicking the offspring allele frequencies that would be present under random fertilization, the equal viability of all gametic genotypes and no natural selection. This first sampling time point was included to account for the possibility that significant allelic changes may have occurred before the first sampling took place – that is, before the trochophore stage, approximately 20h after fertilization. Comparatively, the highest proportion of significant changes in allele frequencies occurred between the time zero ('virtual embryos') sample and the trochophore stage, with the proportion of loci affected ranging from 10% (family J) to 20% (family G). Interestingly, the slight increase observed at the end of our temporal analysis, between day 8 and day 12 larvae, coincides with the age at which the Cawthron hatchery staff have observed significant larval mortalities (personal communication), which motivated the inclusion of this sampling time-point in our study.

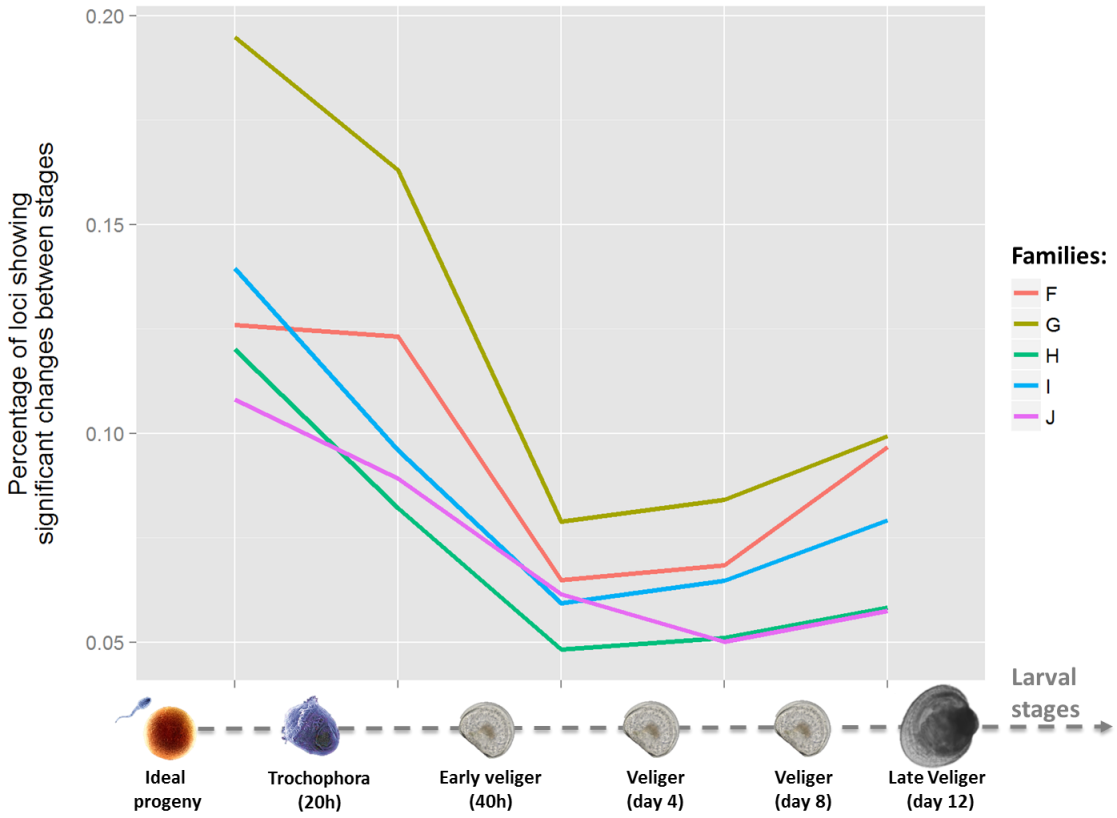


Figure 1. Temporal analysis of changes in allele frequencies in five different Greenshell™ mussel families. In the y-axis the percentage of loci (~100bp in length) analyzed that show significant changes between two consecutive larval stages. In the x-axis, the different time points that were sampled during larval development, including a 'virtual' sample that represented the time-zero time-point. Note that the y-axis does not start from the origin (i.e., zero).

The next stage was to investigate whether there are loci that consistently show distortion (i.e., changes in allele frequency) throughout larval development. For every family, loci were analysed across five larval transitions:

- 'virtual embryo'→ trochophore
- trochophore→ early veliger
- early veliger→ day 4 larvae
- day 4 larvae→ day 8 larvae
- day 8 larvae→ day 12 larvae.

As shown in Table 1, most of the significant changes in allele frequencies affect a single locus at a single larval transition. RAD loci that consistently exhibit changes in allele frequency throughout development are comparatively less (Table 1: All-transitions category).

Table 1. Number of loci that were distorted once, twice or at all larval transitions evaluated in the temporal analysis. The percentage of the total amount of loci showing significant allele changes by family is shown in parenthesis.

	Number of times a loci showed changes in allelic frequencies		
Family ID	**Once**	**Twice**	**All-transitions**
F	261 (46%)	161 (28%)	13 (2%)
G	139 (37%)	133 (35%)	18 (5%)
H	559 (52%)	304 (29%)	16 (2%)
I	379 (50%	211 (28%)	20 (3%)
J	458 (48%)	306 (32%)	20 (2%)

6.3.2 Post-zygotic mutations

To detect mutations arising post-zygotically, novel variants were identified in the pools of mussel offspring at different larval stages. Once reads were mapped to their respective family-specific pseudo-reference a high variation in average read-depth across RAD-loci was

noted; sequencing depth of loci varied according to the type of biological sample, with trochophore samples showing the lowest overall coverage across all family pools. The average read-depth per locus ranged from 26x, in the trochophore pool of family F, to 117x, in the pool of sperm from the sire of family H (Male 19) (Appendix, Figure 3). The lower coverage for certain samples may have affected *de novo* estimations, as mutations appearing at low frequencies in the pool are harder to detect with lower sequencing depths. Nevertheless, the analysis of single stage estimation of *de novo* mutations showed that sample pools with low coverage (i.e., typically the trochophore stage) showed higher or near equal rates of mutations than samples with higher average coverage (Figure 2), suggesting a marginal effect of coverage on the estimates. The frequency of observed putative *de novo* mutations in larval offspring was high and they occurred at all developmental stages (Figure 2). The rate of putative *de novo* mutations by larval stage was calculated by dividing the total number of heterozygous SNPs (i.e., putative mutations) of a pool of larvae over the total number of variants for which both parents are homozygous. When analyzed by larval stage, the day 8 larvae exhibited a higher proportion of putative *de novo* mutations, ranging from 13% to 31%. Lower numbers of putative *de novo* mutations are present in day 12 larvae, with values varying from 3% to 8%. No single larval stage consistently shows the highest or the lowest mutation rates across families. If levels of *de novo* mutations are analyzed by family, then the average frequency of mutations is similar across larval stages, ranging from 12% (in Family F) to 16% (in Families H and I). The Ti/Tv ratio of putative *de novo* mutations was consistent across families (chi^2 test of independence; p-value>0.05) and averaged 0.5, indicating that either a large number of false positives are present in the dataset or that they are true random mutations.

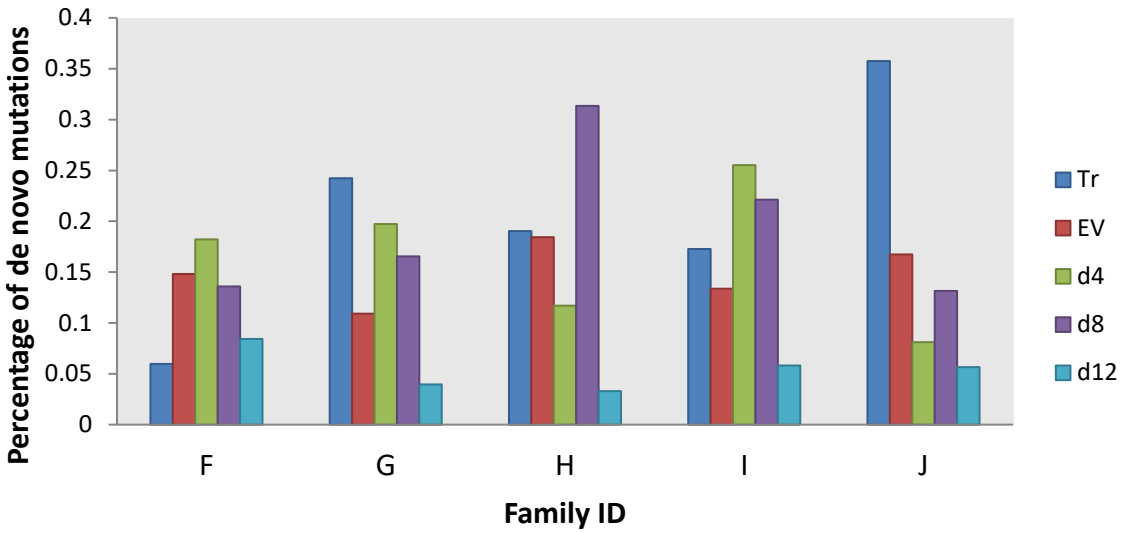


Figure 2. Putative *de novo* mutations detected by larval stage and family. In the x-axis the ID of the family of provenance is shown. In the y-axis the number of *de novo* mutations from the total amount of variants identified in a family are shown, expressed as a percentage. In the legend to the right, the following abbreviation is used for the different larval stages: trochophore (Tr), early veliger (EV), veliger day 4 (d4), veliger day 8 (d8), and veliger day 12 (d12).

Two technical controls of pooled samples were included in the analysis. One had to be removed due to a low number of quality reads. For the remaining sample the genotyping error, defined as the fraction of discordant genotypes between the same DNA sample processed separately (technical replicate), was ~5%. Although this value is within the range of technical variation observed for the RAD-Seq method (Chapter 2 (this thesis) and Mastretta-Yanes et al. (2015)), to be more confident of the putative *de novo* mutations detected in this study, an analysis in which mutations were tracked during development was performed. The rationale behind this analysis is that if a *de novo* mutation appears (post-zygotically) at a certain time in development, then it should either be (i) present in the next larval stage sampled (if the mutation is neutral or sub-lethal) or (ii) absent in any consecutive stages (if the mutation is lethal). In the same manner, if a putative *de novo* mutation appears and disappears at different larval stages, then it is most likely an error or a reflection of the low sensitivity of the methodology to consistently detect alleles throughout development, particularly if these mutations appear at low frequencies. The

earlier trochophore stage was not included in the temporal analysis of mutations because of its comparatively lower quality. The different patterns of presence/absence of alleles during the development of larval offspring were ranked by the number of observations (from the pattern with the greatest number of observations to the lowest), and the top 8 were selected for visualization. As shown in Figure 3, the patterns with a proportionally higher number of observations were those in which a mutation appeared once during development, with no single predominant pattern emerging consistently across families. Patterns in which putative *de novo* mutations emerged in a single larval stage and were absent in the remaining larval stages represented between 67% and 82% of the patterns identified within a family. At a lower frequency (3-15%) we observed different types of patterns that have in common the appearance of a mutation in an early larval stage, the maintenance of the mutation in the next larval stage, but then the disappearance of the novel allele in the last sampling time point (i.e., day 12 larval pools). It is noteworthy that none of the most abundant patterns fitted with what would have been expected in the presence of significant cross-contamination with exogenous DNA.

One reason that may explain the ephemeral nature of the putative *de novo* mutations observed when the fate of mutations is tracked during larval development is technical error. The fact that we observe patterns in which a mutation appears in an early larval stage, disappears in the next temporal stage and then re-appears in the following one may be an indication of a certain degree of technical noise that is impeding the correct sampling of the novel alleles. Furthermore, given the majority of presence/absence patterns derived from larval pools are the ones where mutations are detected at a single particular larval stage but are not present either earlier or later in development, raises concerns regarding the sensitivity of the experiment to accurately detect mutations in pools of mussel larvae. To evaluate the sensitivity of the methodology, the frequency of the *de novo* allele in each larval pool was calculated by family. As shown in Figure 4, putative post-zygotic *de novo* mutations occur at a low frequency (<0.07), showing a right-skewed distribution with a mode at either 0.02 or 0.03 depending on the family. Therefore, coverages per locus of 500x would be required to accurately identify variants present at 2.5% in a larval pool (Stead et al., 2013). At the moderate coverage depths that sample pools were sequenced in this study (maximum effective average of 117x, although the experiment was designed to achieve 200x), correct assignment of the absence or presence of novel alleles was limited.

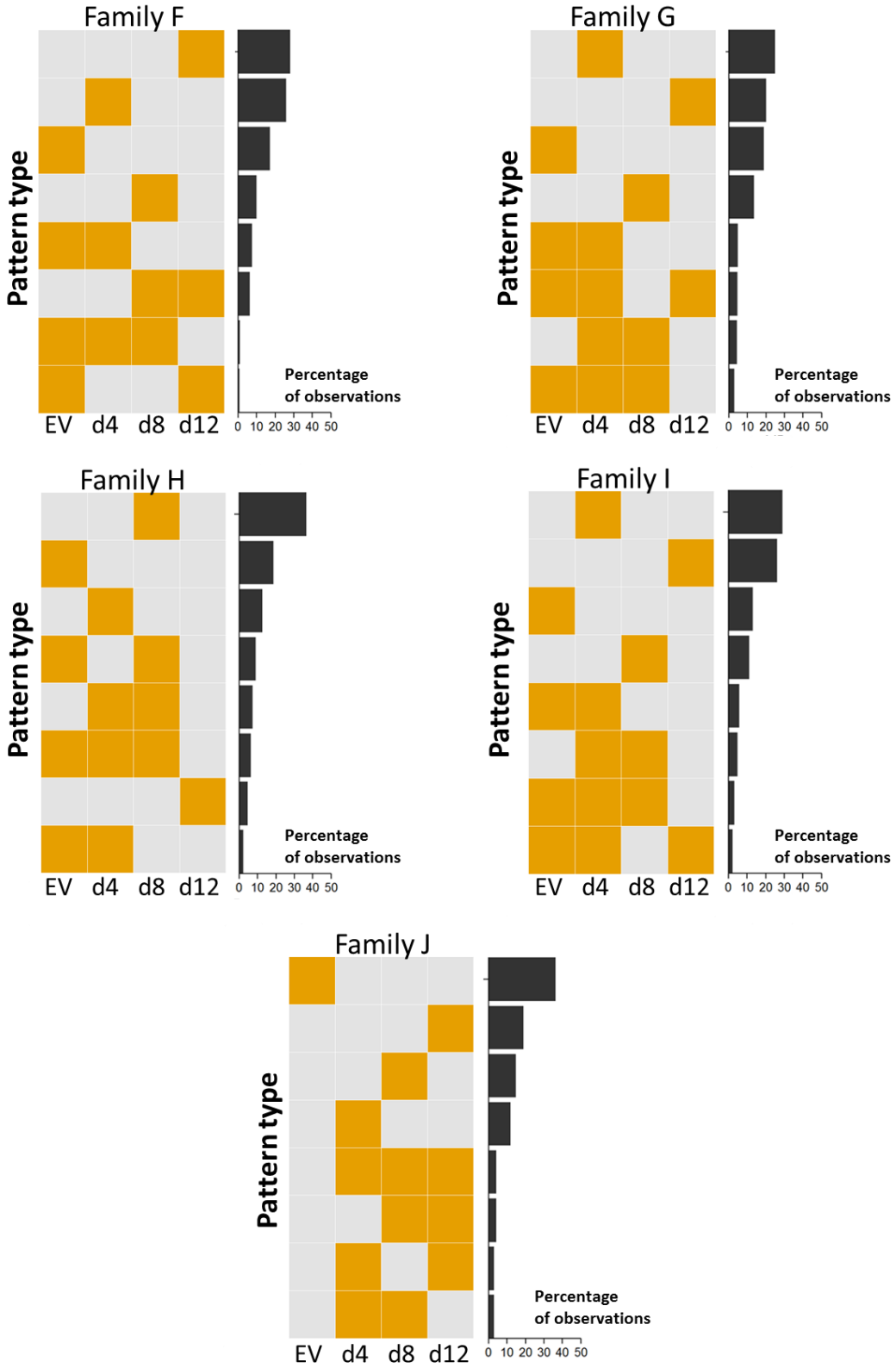


Figure 3: Patterns of presence/absence of *de novo* mutations at different stages of larval development. In the x-axes the different larval stages coded as follows: early veliger (EV),

day 4 veliger larvae (d4), day 8 veliger larvae (d8), and day 12 veliger larvae (d12). In the y-axes (to the left), the most frequent patterns of absence/presence of *de novo* mutations during development (in decreasing order). The patterns are interpreted as follows: for all loci detected within a family a genotype comparison across different larval stages was made; if the genotype of a larval pool was homozygous then the square is grey, whereas if the inferred genotype was heterozygous (indicating a putative *de novo* mutation, as both parents were homozygous for the locus) an orange square is depicted. In the y-axes (to the right), the percentage of observations that show a specific pattern.

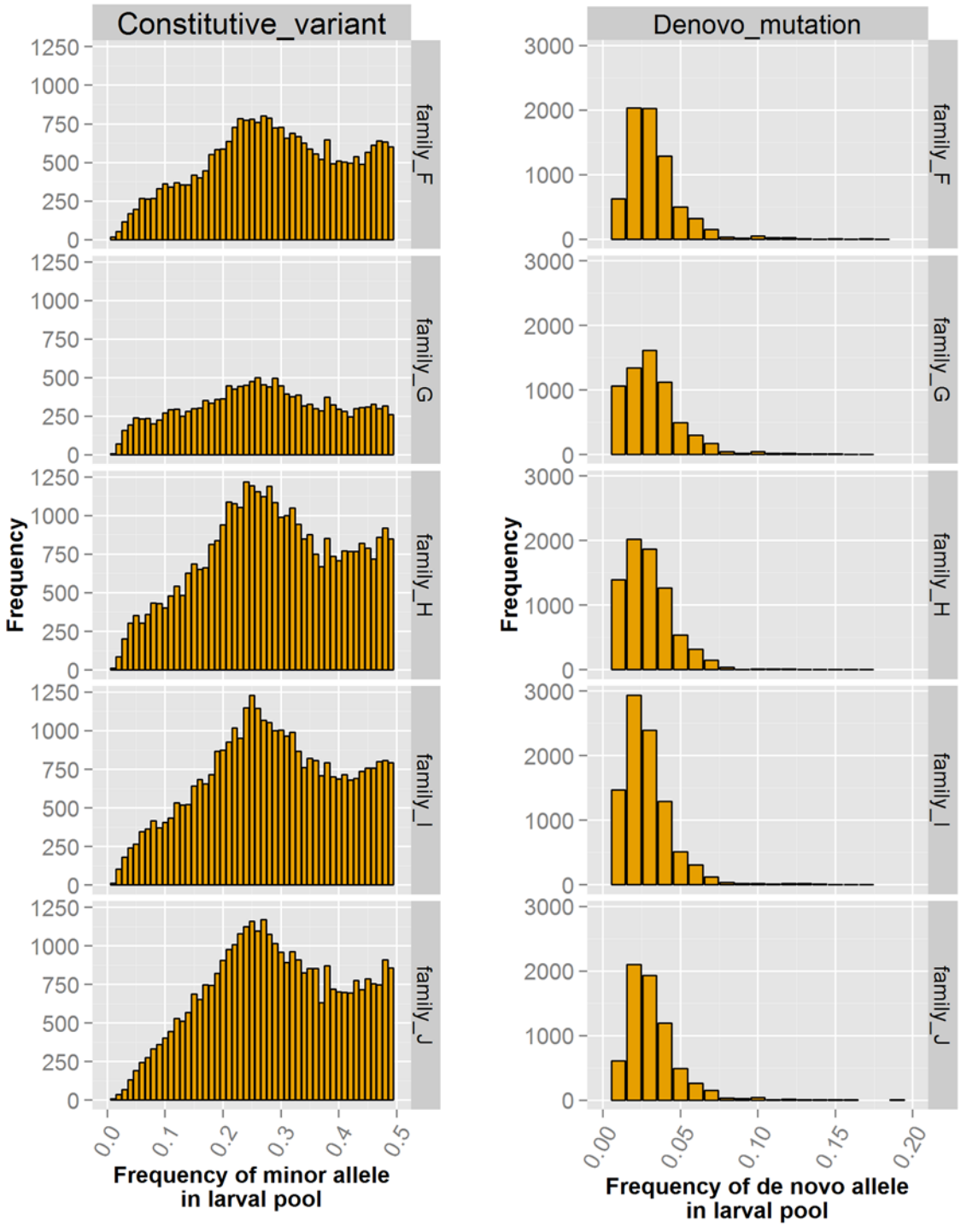


Figure 4. Comparison of the minor allele frequency of SNP markers that are constitutive of a family (i.e., a heterozygous SNP is observed across all larval stages) (left) and the frequency of a *de novo* allele that is detected at least once in any larval stage (right).

6.3.3 Transmission ratio distortion

Meiotic drive is characterized by the differential formation or success of gametes, which results in heterozygous markers in the parents showing departures from the expected 1:1 ratio. We evaluated the possibility of TRD in male and female mussels by estimating the allele ratios of heterozygous SNP markers in the parents and their respective gametes. Under optimal (technical) conditions the allele ratio of a heterozygous marker should be similar in an individual parent compared to their gametes. For each SNP marker, it was tested whether the allele ratio of a single parent deviated significantly from the allele ratio determined for its pool of gametes. Our results show that the markers showing evidence of TRD (Figure 4), were not significant after Bonferroni correction. Therefore, we find no significant evidence of TRD in these mussels. The allele ratios of parent vs. gamete showed a higher consistency in males compared to females, given their homogeneous distribution across the diagonal. Females, on the other hand, show a greater oscillation in the relationship between the two allelic ratios, possibly reflecting the lower DNA quality of female gametes compared to the male counterpart (Appendix, Section 1: Figure 3).

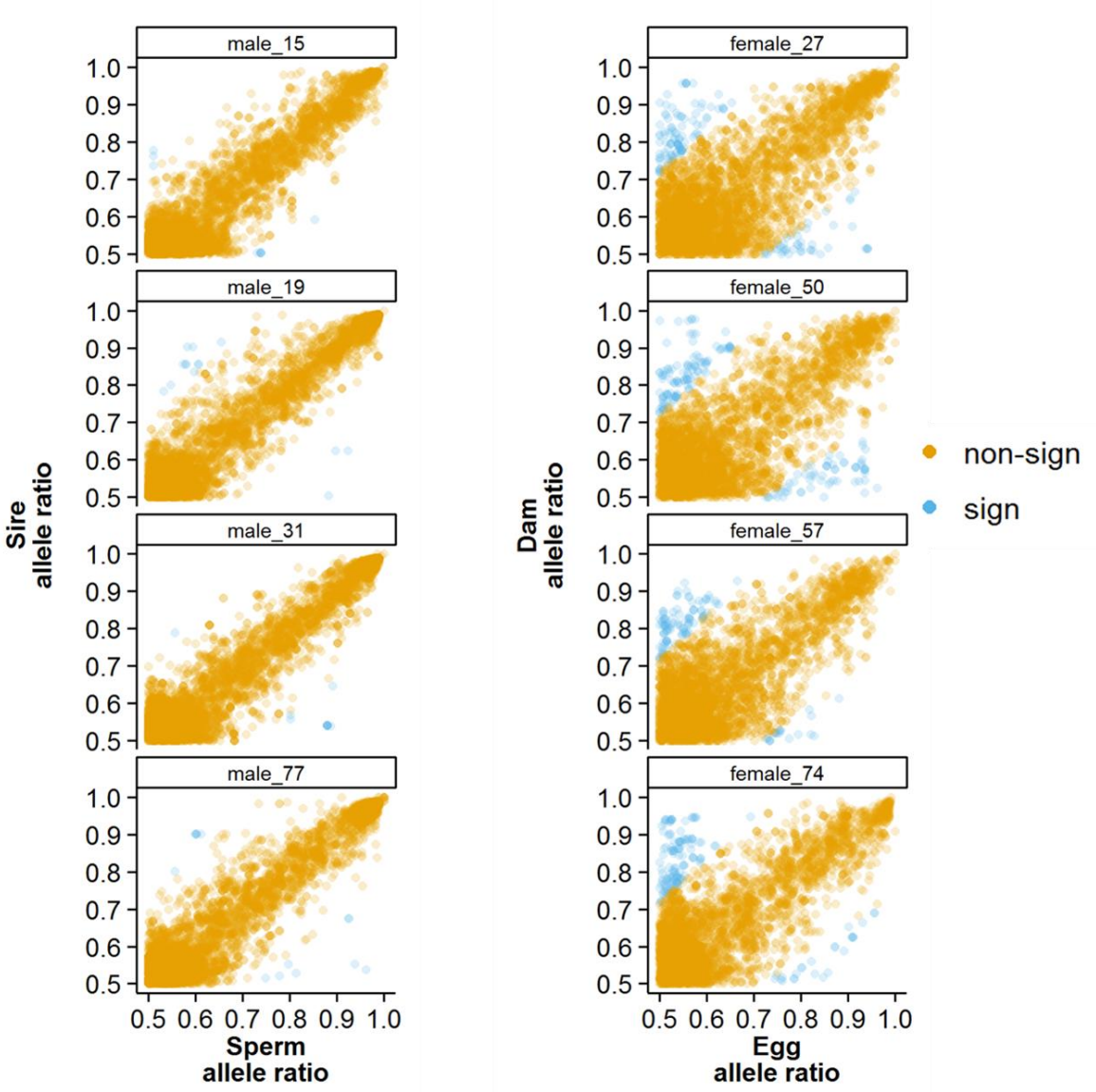


Figure 7. Transmission ratio distortion (TRD) evaluation on male and female individuals used as sires and dams in this study.

6.3.4 Germline *de novo* mutations

Table 2 shows the number of putative *de novo* mutations detected in the gametes (*parent-gametes* dataset) of the four male and four female mussels used to create the families analysed in this study. A high number of novel alleles were identified in both male and female Greenshell™ gametes. Moreover, mutation rates are surprisingly high considering that <1% of the Greenshell™ mussel genome was examined. Overall, the number of mutations was higher in females than in males, with an average of 222 in females, 28.5 in males, and a total average across gamete pools of 125 *de novo* mutations.

The fact that consistently a higher overall number of *de novo* mutations were detected in all females compared to male individuals is an argument against these variants being false positives generated by systematic artifacts. The main source of false positives is errors in the alignment of short reads to a reference genome (Koboldt, 2013). If this was the case in our data, male and female individuals would have been equally affected, showing no bias to a certain gender. The range of mutations was different between males and females, with females showing a wider range of mutations compared to males (t-test, p-value=0.0064 for single nucleotide mutations; p-value=0.014 for LOH events). The difference between the male with the highest versus the lowest number of mutations was 2.2 fold, compared to a difference of 3.1 in females. In addition, a higher frequency of LOH events is present in females. Putative LOH events tended to be lower in frequency compared to putative *de novo* mutations in females, whereas in males they were higher. It is noteworthy that the greater proportion of *de novo* mutations in female gametes is evidence against the possibility that the generally lower quality of the DNA extraction of eggs biased our results. As a consequence of low quality, a higher risk of not sampling both alleles would be expected. This would lead to an artificial decrease in the number of *de novo* mutations, which is not consistent with the data presented in Table 2, indicating a minor effect of lower DNA quality in these incipient stages of analysis of the *de novo* mutation in mussel gametes.

After detecting a set of putative *de novo* mutations in the gametes of male and female mussels, we sought to determine if they got transmitted to the first larval stage sampled (trochophore → ~20 hpf). The numbers of *de novo* mutations detected in the offspring are shown in Table 3. Following the trend observed in the gametes, more putative *de novo* mutations are transmitted from the female parent compared to the paternal line.

Table 2. A higher number of *de novo* mutations and LOH events were detected in the germline of the Greenshell™ mussel dams compared to sires. Results are presented for the pairwise comparison of each parent (sire or dam) with their respective gamete pool.

Sample ID	Family ID	Number of *de novo* mutations	Number of LOH events	Total N° of SNPs
Female 27	H	98	135	9,863
Female 50	G	276	154	9,644
Female 57	F & J	203	113	9,889
Female 74	I	311	164	9,472
Male 15	G	38	120	9,650
Male 19	H & J	35	52	8,563
Male 31	I	17	21	9,141
Male 77	F	24	29	9,023

Table 3. Frequency of germline *de novo* mutations of maternal and paternal origin that were also detected in the first larval stage sampled.

	Offspring - age 20h	
Family ID	**Maternal origin**	**Paternal origin**
F	18	0
G	26	13
H	16	0
I	29	3
J	18	0

By knowing the frequency of the *de novo* allele in the gamete pool and the read-depth of the locus, a general view of the mutation frequency may be obtained. Comparatively, the frequency of *de novo* mutations is higher in female than in male gametes (Figure 5). In eggs,

the *de novo* allele frequency ranged from 0.05 to 0.7, with a mean of 0.11 and a mode of 0.06. In sperm the range was narrower, extending from 0.05 to 0.34, with a mean of 0.09 and a mode of 0.06. The detection of *de novo* alleles at a lower frequency was impeded due to the coverage thresholds applied: minimum of 3 reads to call an allele at a maximum of 100x coverage per locus. As shown in Figure 6, at lower read-depths (<50x) both male and female gametes exhibit *de novo* mutations at frequencies higher than 0.05. Nevertheless, these high mutation frequencies are not detected at higher coverages (>50x). This unequal dispersion of the data is most likely the result of an imbalanced allelic representation at low coverages. The correlation coefficient between the (i) frequency of *de novo* alleles and (ii) locus read-depth for the pool of eggs was -0.33, indicating a weak negative correlation between variables. On the other hand, in the pools of Greenshell™ mussel sperm, only a few *de novo* mutations are detected at read-depths <50x, with the majority identified at coverages between 50x-100x. For the pools of sperm, a low correlation between frequency of *de novo* alleles and locus read-depth was found (Pearson correlation coefficient=-0.19). Since novel alleles in both the male and female gamete pools asymptote at ~0.05, it can be concluded that putative germline mutations occur at low frequencies. Thus they can solely be reliably detected at high coverages. Considering that a minimum of 3 reads was required to call an allele, any locus should exhibit a sequencing depth of at least 60x for a *de novo* allele to be detected at a frequency of 0.05. At the moderate coverages that were observed in the egg pools (Appendix, Section 1: Figure 3), it is likely that there is a higher rate of false-negatives in the egg dataset.

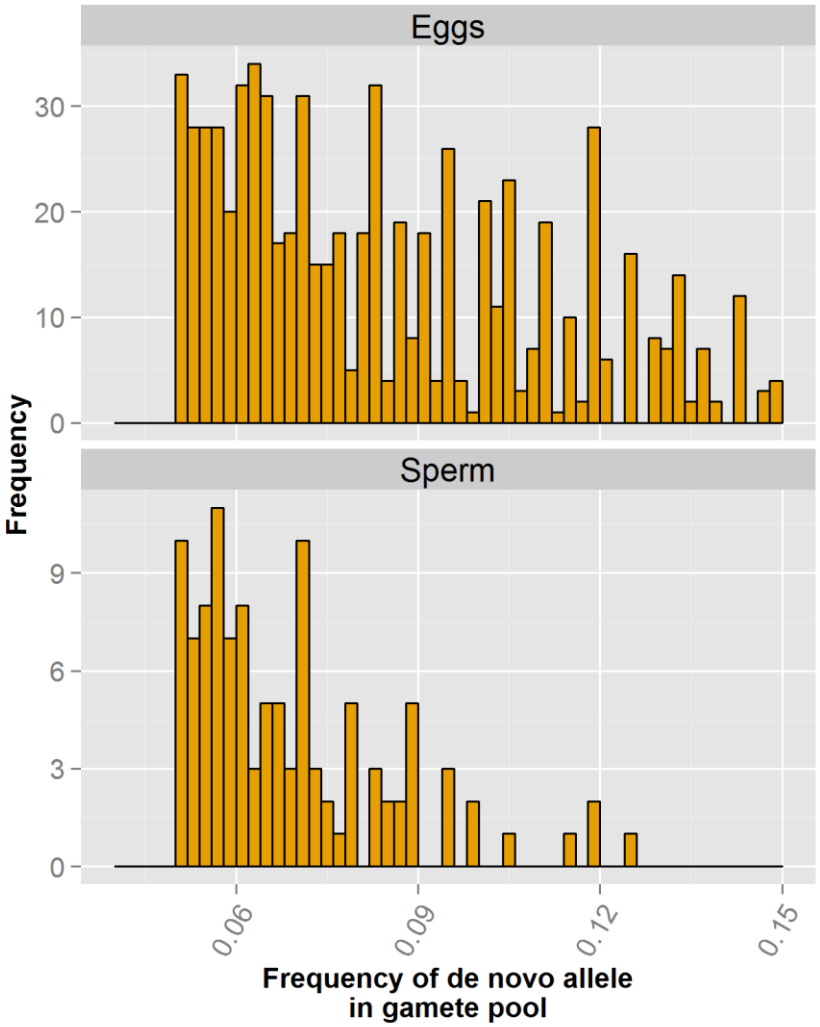


Figure 5. Comparison of the frequency of *de novo* alleles between pools of eggs (top) and sperm (bottom). In the x-axes the minor allele frequency of the *de novo* allele, and in the y-axes the number of observations (frequency) for each category. Only minor allele frequency values between 0.05 and 0.15 are shown.

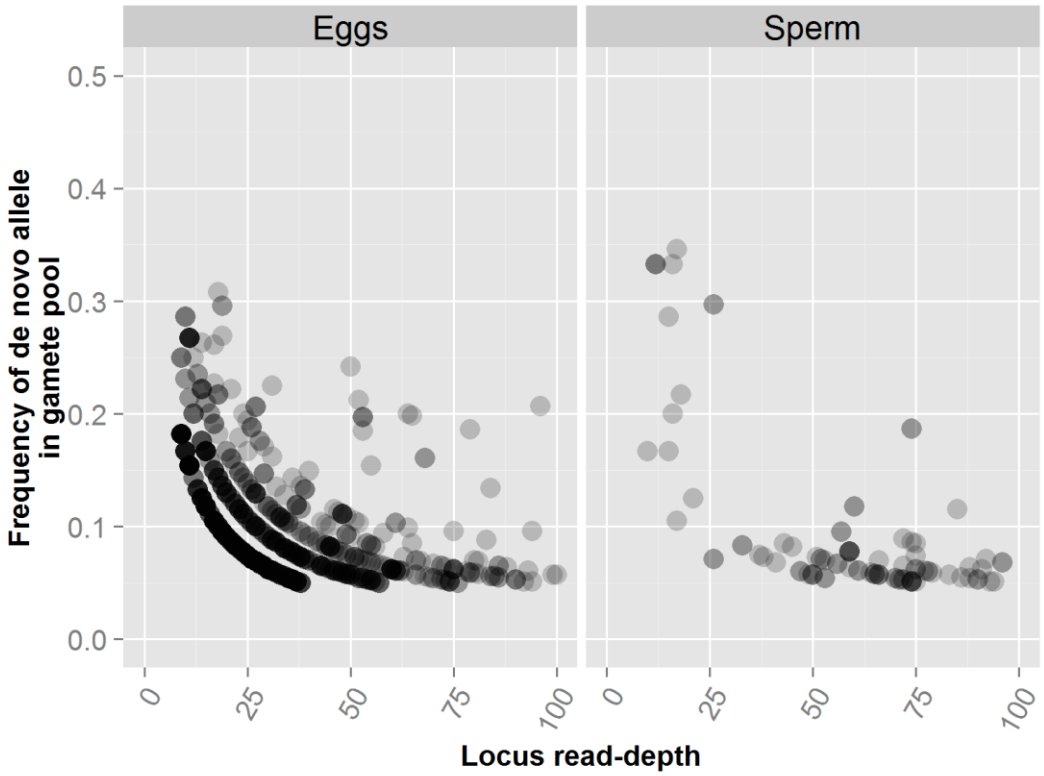


Figure 6. The frequency of putative *de novo* alleles asymptotes at ~0.5 at higher coverages in eggs (left side) and sperm (right side). In the x-axes the read-depth of a locus that shows a novel allele. The y-axis shows the frequency of the *de novo* allele in the pool of gametes.

6.4 Discussion

The RAD sequencing of five Greenshell™ mussel families identified a significant number of putative *de novo* alleles appearing in a single larval generation, overall representing around 15% of the discovered single nucleotide variation at the RAD loci. These putative mutations were detected mostly in the germline of female mussels and in the larval progeny, suggesting that two independent sources – gametogenesis and post-zygotic events – are contributing to the short-term genetic makeup of early developmental stages of a bivalve species.

6.4.1 Sources of *de novo* mutations in mussels

Post-zygotic mutations have not been studied as potential sources of genetic variation in bivalve species. Previous genetic studies in mussels – although not formally searching for mutations – however have consistently reported unexpected genotypes in the adult offspring of controlled pair crosses, which did not fit with the presence of null alleles (2%-8% in Reece et al. (2004); 5% in McGoldrick and Hedgecock (1997)). Due to the generally low frequency of these individuals, they were treated as contaminants. Nevertheless, it is reasonable to suggest, particularly in view of the rates of mutation detected in the Greenshell™ mussel families, that these 'contaminants' may have been the result of *de novo* mutations. This may also suggest that the high mutation rate predicted for bivalves possibly creates such mutational burdens that unfit genotypes get removed by natural selection early during development (Plough, 2016), before sampling takes place. Consequently, in bivalves, the main lifecycle stage in which to search for *de novo* mutations is the larval stage.

By comparing the RAD-Seq data of parents versus their progeny at different larval stages, not only was a high level of putative mutation identified (from ~200 to 6,000 novel alleles) but the fate of these mutations was tracked through the larval life cycle. The Greenshell™ mussel larvae raised in this study settled at day 15, therefore, from fertilization to day 12 (the last time-point sampled) approximately 80% of the larval cycle was evaluated. De novo mutations (as defined by comparing pools of larvae to parents) were detected at all the pools of larval stages sampled: trochophore and veligers sampled at day 4, day 8 and day 12. The levels of novel alleles were significantly higher than the estimated genotyping error

of 0.4 % for the Illumina platforms (Quail et al., 2012). Thus, there is a low risk that mutations are being confounded by sequencing errors. The amount of *de novo* mutations per larval stage, counted across replicates and families, revealed no pattern that may indicate that a certain larval stage (or age) is more susceptible to mutations (Figure 2). For instance, the stage at which more *de novo* mutations were detected in family J was the trochophore stage, comprising ~36% of all genetic polymorphisms identified in this family. In comparison, in family H the majority of post-zygotic mutations appeared in the veliger larvae 8 days after fertilization, which represented ~31% of the *de novo* variants within the family. This lower level of mutation may indicate that there is a biological window at which larvae are more vulnerable to mutational events. When the fate of mutations was tracked during development, as a means of assessing the nature of mutations (neutral, lethal, sub-lethal), in most occasions mutations appeared at a certain larval stage but disappeared in the next sampled larval stage. In this context, it is difficult to imagine why we did not observe, at proportionally higher frequencies, patterns (Figure 3) that would fit with the arising of a neutral mutation – that is, a pattern that consistently showed the presence of the mutation in all consecutive stages after its appearance in the larval pool. A limiting factor in the use of pooling strategies is the sensitivity of the assay. By sensitivity, we mean the probability to detect a mutation given that the mutation is present in some member(s) of the DNA pool. Although the average marker coverage in the larval pools was ~90x, a specific locus would have to be consistently sequenced at read-depths >60x in all larval stages in order to consistently detect *de novo* alleles at frequencies <0.05, which is highly unlikely. Due to this technical limitation, the true nature – whether neutral, lethal, or sub-lethal – of these putative post-zygotic mutations remains to be assessed. The fact that mussel larvae exhibit these high amounts of mutations under hatchery conditions, in the absence of genotoxic substances that may have unexpectedly induced mutational events (Bouilly et al., 2003), indicate they are an intrinsic property of the biology of bivalves, as predicted (Plough, 2016). Given bivalves experience significant chromosomal abnormalities and can tolerate cell chromosome number variations ranging between 5 and 30% (Leitão et al., 2001, Thiriot-Quievreux, 2002), it is likely that post-zygotic point mutations are part of a wider spectrum of *de novo* polymorphisms that are actively affecting the genomic integrity of bivalves.

Another source of *de novo* mutations evaluated in this study was gametogenesis. The RAD-Seq data of male and female mussels were compared with their respective sequenced pool

of gametes. The vast majority of germline *de novo* alleles are maternally derived (~89 % in our study). A higher *de novo* mutation rate for females is contrary to the expectation of sperm carrying a higher mutational load due to a comparatively higher number of germline cell divisions (Kahn and Quinn (1999), although see (Hurst and Ellegren, 1998)). A single female mussel can spawn between 5 and 12 million eggs annually. Hence, they also must undergo the high number of germline cell divisions experienced by males. The fundamental difference observed between the mutation rates of sexes suggest distinct mutational mechanisms are operating in mussel spermatogenesis and oogenesis. Evidence in humans indicates that the signatures of mutation depend on the parent of origin, as a result of differences in the biology of male and female gametogenesis. For example, paternally derived mutations are enriched for T→G and C→A substitutions, whereas maternally derived ones contain a higher amount of C→T changes. Additionally, genomic location and nucleotide context also revealed parent-of-origin signatures (Goldmann et al., 2016). Several attempts have been made to uncover biological processes of mutations based on genome signatures, particularly in cancer research (Greenman et al., 2007, Alexandrov et al., 2013). The application of a similar approach may aid in the understanding of the mutational processes underlying the difference between male and female mussel mutation rates observed in this study. In addition to gender differences in rates of mutations, we observed a significant variation among individuals within sexes. A 2.2-fold difference was found between the male individual with the highest versus the lowest number of mutations. For females this difference spanned 3.1 orders of magnitude across individuals. This natural inter-individual variation raises the question if there are genetic (heritable) factors that may be controlling the rate of mutation of DNA. In this study, the two half-sib families that shared the same dam produced offspring with the most dissimilar mutational patterns observed across larval development (Figure 3: families F and J). However, with the small number of families analyzed herein, it is not possible to derive any reliable conclusion regarding the role of genetic factors in the mutation rate of mussels. Nevertheless, others have reported that chromosomal loss in bivalves is not random (de Sousa et al., 2012) and that aneuploidy (i.e., presence of an abnormal number of chromosomes in a cell) has a genetic basis, given levels of aneuploidy have been observed to be inherited from parents to offspring (Leitão et al., 2001). The results of the current study demonstrate that a fraction (~12%) of the *de novo* mutations detected in male and female gametes were present in the first sampled time-point, at the trochophore stage ~20h after fertilization.

This implies that they are actively contributing to the genetic make-up of the larval population. However, larval development represents a complex biological process, characterized by dramatic morphological, physiological, and behavioral changes. Moreover, during metamorphosis, the complete reorganization of the body plan is finely orchestrated (Wygoda et al., 2014), suggesting that the manifestation of mutations at this stage would be highly detrimental. Whether maternally or paternally inherited mutations can survive this complex larval cycle, produce viable and fertile individuals, and get transmitted to the next generation, remains to be addressed.

6.4.2 Temporal changes in allele frequencies

The larval offspring of five mussel families were monitored for changes in allele frequencies during development. A similar pattern was observed across families, suggesting a consistent process is taking place (Figure 1). The number of loci showing significant changes in allele frequencies was proportionally higher at the earliest larval stages, steadily decreasing during development, with a final slight increase between the last larval transition examined (day 8→ day 12). Previous literature indicates that segregation distortions (or deviations from Mendelian inheritance) occur at a higher frequency around the time of metamorphosis, when many genes are likely to be activated for the first time (Launey and Hedgecock, 2001). We report that significant distortions also occur earlier in development, probably at the embryonic stage. Our results show that changes in allele frequencies tend to occur at significant magnitudes at the time window between the time zero ('virtual') sample and the early veliger larva (in our samples represented by day 4). Additionally, we find that loci associated with changes in allele frequencies were distorted mainly once during development, instead of consistently being responsible for distortions throughout larval development. This may suggest either (i) that distinct genomic regions are temporally expressed during development, and therefore differentially selected; or (ii) post-zygotic *de novo* mutations are (partially) driving local temporal loci effects. An approach to discern between these two possibilities would be to examine whether the same loci are affected during the same larval transition in other mussel families, which if true would reflect a biological effect. However, given we aligned the sample reads to family-specific pseudo references, we are not able to fully test this hypothesis.

6.4.3 Transmission ratio distortion

The frequent observation that markers in bivalves typically show segregation distortions (or deviations from Mendelian proportions) raises the possibility that pre-zygotic processes may be playing a role. Although it is not the purpose of this Chapter to examine potential sources of segregation distortion, we made use of the data generated in this study to evaluate a different aspect of segregation distortion that was not covered in the previous Chapters. We utilized the RAD-Seq data from male and female parents and a pool of their gametes to detect consistent deviation from the 1:1 expectation of a heterozygous marker. Each allele from an individual should be transmitted with equal probability to the offspring. Hence, a pool of sequenced gametes (1 gamete=1 haploid cell=1 chromosome pair=1 sequencing read) should contain half of the reads matching one allele and half to the other allele. To correct for potential biases that may arise due to sample type or the sequencing process, for each SNP marker we compared the allelic ratios of the parent vs. gametes. As both ratios should be similar at near optimal conditions, we fitted the allelic ratio (parent: gametes) to a 1:1 expectation. Our results show no evidence of TRD in male or female individuals. We therefore conclude that segregation distortions arise in bivalves through a different mechanism than the preferential transmission of one allele from either parent.

6.4.4 Technical limitations

Two major potential sources of technical noise were present in this study. First, a considerable variability in DNA quality across samples was observed. This variability was associated with the type of biological material analyzed. DNA extracted from eggs (oocytes), and trochophore larvae represented the lowest quality samples. As can be observed in the coverage per RAD-loci graph across samples (Appendix, Section 1: Figure 2), ~90% of the loci in these samples have a coverage <100x. A maximum of 300x coverage per loci was allowed in our analysis; this upper limit favoured the detection of *de novo* mutations in samples of better quality and of consequently higher coverage. This difference in average read depth by sample type may have affected the ability to detect *de novo* mutations in pools of eggs or trochophore. Nevertheless, the impact of this limitation will depend on the frequency of *de novo* mutations. In this study we observe a high frequency of mutations in the pools of both eggs and trochophore larvae, indicating that an allele

sampling bias, if present, has a minor impact on the comparatively high rates of mutation detected. Another source of potential bias in our analysis may have been generated by the RAD-Seq technique. The genotype discordance rate is defined herein as the number of genotype calls that differ between a pair of replicate samples (same DNA) divided by the total number of non-missing SNPs. The genotype discordance estimated from technical controls of larval pools was around 5%. This high value of genotype discordance is most probably caused by the sub-optimal sampling of all the alleles present in the pool of larvae. I must be noted, however, that under the parameter values that we used, which were set to detect alleles appearing at a low frequency, a higher discordance rate is expected. At present, without an independent technique to validate our findings, false-positives and false-negatives co-exist with true *de novo* variants in our dataset. The most favorable situation to validate mutations would be to sequence single larvae (an attempt that we make in Appendix, Section 2). However, the main limitation of this approach is the high effort involved to detect low-frequency mutations, as a high number of individuals would have to be genotyped. Nevertheless, these confirmatory experiments are necessary to validate some of the results described herein.

6.4.5 Evolutionary implications

Mutations are the ultimate source of novel genetic variation. Mutations typically arise spontaneously at a low frequency (e.g., ~1.2×10^{-8} substitutions per generation in humans; (Campbell et al., 2012)), and are therefore thought to exert marginal effects on the genetics of populations. However, the high rates of *de novo* mutations observed in this study suggest they may play a major role in the short-term response to selection of bivalves. Of most relevance to our findings are the arguments exposed by William's in his theoretical work on the evolution of sexual reproduction (Williams, 1975). The author proposed that as a consequence of early selection in a fluctuating environment, high-fecundity organisms would experience extensive genetic change in a single generation. Although William's utilized bivalves as a model to describe the advantage of sexual reproduction (Elm-Oyster model), he lucidly extracts the essential features of fecund organisms, providing a link between life history traits and certain defining genetic properties of an organism. Indeed, the parallels between the population biology of long-lived, highly fecund plants and bivalve mollusks are remarkable and include their high fecundity, high propagule dispersal, and the

fact they are sedentary most of their lifespan. At high fecundities but low population sizes, as observed previously for wild bivalve populations (Gosselin and Qian, 1997), strong selection pressures must be taking place. William's argued for a relationship between high fecundity and selection at early stages, suggesting that part of these mortalities were associated with a high genetic load (see, for a review, (Plough, 2016)). Genetic load is understood as the difference between a particular genotype in populations and the theoretically fittest genotype. With a high genetic load, the proportion of the deaths which arise from variations in fitness is higher. Direct evidence of high genetic load in plants has been provided, for example, by counting abortions (Wiens et al., 1987, Kärkkäinen et al., 1999). For marine bivalves, on the other hand, the empirical demonstration of genotype-dependant mortalities remains elusive, mainly because of the complexities derived from the experimental design (e.g., the sampling of inviable individuals). Nevertheless, indirect evidence of high genotype-dependant mortalities at larval stages has been provided based on a retrospective analysis of SD (Plough and Hedgecock, 2011, Plough, 2012). If genetic load is the negative aspect of a high mutation rate, then faster adaptation may be the positive. Although most mutations are neutral or deleterious, some will be advantageous under local heterogeneous environments, which may be opportunistically exploited by certain genotypes. A relevant issue under this scenario would be how bivalves maintain a high mutation rate, as it should be, by definition, unsustainable in the long-term – a mutator allele (i.e., an allele that increases the mutation rate) would be itself eventually disrupted by mutations. Indeed, several key questions remain to be addressed, which in the face of remarkably high levels of putative *de novo* mutations detected in this study, may require the re-evaluation of the concepts of inheritance, population genetics, and evolution applied for marine bivalves.

In summary, we report a high rate of *de novo* mutations in a bivalve species, the Greenshell™ mussel. By analyzing genome-wide patterns of genetic variation of different life-history stages of the mussel life-cycle – parents (adults), their gametes, and larval offspring at different ages – we detected *de novo* mutations originating from two sources, during gametogenesis or post-zygotically in larvae. To the best of our knowledge, this is the first time RAD-Seq is utilized for uncovering fundamental biological aspects of larval genetics**.** The increasing accessibility to NGS technologies will serve to validate the data presented here and provide estimates of mutation rates in other bivalve species, yielding

further insight into the evolutionary significance of the extreme levels of polymorphisms present in marine bivalves.

6.5 References

ADAMS, S. L., TERVIT, H. R., MCGOWAN, L. T., SMITH, J. F., ROBERTS, R. D., SALINAS-FLORES, L., GALE, S. L., WEBB, S. C., MULLEN, S. F. & CRITSER, J. K. 2009. Towards cryopreservation of Greenshell™ mussel (Perna canaliculus) oocytes. *Cryobiology,* 58**,** 69-74.

ALEXANDROV, L. B., NIK-ZAINAL, S., WEDGE, D. C., APARICIO, S. A., BEHJATI, S. & BIANKIN, A. V. 2013. Signatures of mutational processes in human cancer. *Nature,* 500, 415-421.

ANDREWS, K. R., GOOD, J. M., MILLER, M. R., LUIKART, G. & HOHENLOHE, P. A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics,* 17**,** 81-92.

ARANA, M. E. & KUNKEL, T. A. 2010. Mutator phenotypes due to DNA replication infidelity. *Seminars in cancer biology,* 20**,** 304-311.

BIERNE, N., LAUNEY, S., NACIRI-GRAVEN, Y. & BONHOMME, F. 1998. Early Effect of Inbreeding as Revealed by Microsatellite Analyses on Ostrea edulis Larvae. *Genetics,* 148**,** 1893-1906.

BUSHNELL, B. 2016. BBMap short read aligner. *University of California, Berkeley, California. URL http://sourceforge. net/projects/bbmap*.

CAMPBELL, C. D., CHONG, J. X., MALIG, M., KO, A., DUMONT, B. L., HAN, L., VIVES, L., O'ROAK, B. J., SUDMANT, P. H., SHENDURE, J., ABNEY, M., OBER, C. & EICHLER, E. E. 2012. Estimating the human mutation rate using autozygosity in a founder population. *Nature Genetics,* 44**,** 1277-1281.

CATCHEN, J. M., AMORES, A., HOHENLOHE, P., CRESKO, W. & POSTLETHWAIT, J. H. 2011. Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes|Genomes|Genetics,* 1**,** 171-182.

DE SOUSA, J. T., JOAQUIM, S., MATIAS, D., BEN-HAMADOU, R. & LEITÃO, A. 2012. Evidence of non-random chromosome loss in bivalves: Differential chromosomal susceptibility in aneuploid metaphases of Crassostrea angulata (Ostreidae) and Ruditapes decussatus (Veneridae). *Aquaculture,* 344**,** 239-241.

ELLEGREN, H. & GALTIER, N. 2016. Determinants of genetic diversity. *Nature Reviews Genetics,* 17**,** 422-433.

FU, L., NIU, B., ZHU, Z., WU, S. & LI, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics,* 28**,** 3150-3152.

GOLDMANN, J. M., WONG, W. S. W., PINELLI, M., FARRAH, T., BODIAN, D. & STITTRICH, A. B. 2016. Parent-of-origin-specific signatures of de novo mutations. *Nature Genetics,* 48, 935-939.

GOSSELIN, L. A. & QIAN, P.-Y. 1997. Juvenile mortality in benthic marine invertebrates. *Marine Ecology Progress Series,* 146**,** 265-282.

GREENMAN, C., STEPHENS, P., SMITH, R., DALGLIESH, G. L., HUNTER, C., BIGNELL, G., DAVIES, H., TEAGUE, J., BUTLER, A., STEVENS, C., EDKINS, S., O'MEARA, S., VASTRIK, I., SCHMIDT, E. E., AVIS, T., BARTHORPE, S., BHAMRA, G., BUCK, G., CHOUDHURY, B., CLEMENTS, J., COLE, J., DICKS, E., FORBES, S., GRAY, K., HALLIDAY, K., HARRISON, R., HILLS, K., HINTON, J., JENKINSON, A., JONES, D., MENZIES, A., MIRONENKO, T., PERRY, J., RAINE, K., RICHARDSON, D., SHEPHERD, R., SMALL, A., TOFTS, C., VARIAN, J., WEBB, T., WEST, S., WIDAA, S., YATES, A., CAHILL, D. P., LOUIS, D. N., GOLDSTRAW, P., NICHOLSON, A. G., BRASSEUR, F., LOOIJENGA, L., WEBER, B. L., CHIEW, Y.-E., DEFAZIO, A., GREAVES, M. F., GREEN, A. R., CAMPBELL, P., BIRNEY, E., EASTON, D. F., CHENEVIX-TRENCH, G., TAN, M.-H., KHOO, S. K., TEH, B. T., YUEN, S.

T., LEUNG, S. Y., WOOSTER, R., FUTREAL, P. A. & STRATTON, M. R. 2007. Patterns of somatic mutation in human cancer genomes. *Nature,* 446**,** 153-158.
HURST, L. D. & ELLEGREN, H. 1998. Sex biases in the mutation rate. *Trends in Genetics,* 14**,** 446-52.
KÄRKKÄINEN, K., SAVOLAINEN, O. & KOSKI, V. 1999. Why do plants abort so many developing seeds: bad offspring or bad maternal genotypes? *Evolutionary Ecology,* 13**,** 305-317.
KOBOLDT, D. C. 2013. Using VarScan 2 for Germline variant calling and somatic mutation detection. *Current Protocols in Bioinformatics,* 44, 15.4.1-15.4.17.
KOBOLDT, D. C., ZHANG, Q., LARSON, D. E., SHEN, D., MCLELLAN, M. D., LIN, L., MILLER, C. A., MARDIS, E. R., DING, L. & WILSON, R. K. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research,* 22**,** 568-76.
KOFLER, R., OROZCO-TERWENGEL, P., DE MAIO, N., PANDEY, R. V., NOLTE, V., FUTSCHIK, A., KOSIOL, C. & SCHLÖTTERER, C. 2011. PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLoS ONE,* 6**,** e15925.
LAUNEY, S. & HEDGECOCK, D. 2001. High genetic load in the Pacific oyster Crassostrea gigas. *Genetics,* 159**,** 255-265.
LEITÃO, A., BOUDRY, P., MCCOMBIE, H., GÉRARD, A. & THIRIOT-QUIÉVREUX, C. 2001. Experimental evidence for a genetic basis to differences in aneuploidy in the Pacific oyster (Crassostrea gigas). *Aquatic Living Resources,* 14**,** 233-237.
LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* 25**,** 1754-1760.
LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.
MACAVOY, E. S., WOOD, A. R. & GARDNER, J. P. A. 2008. Development and evaluation of microsatellite markers for identification of individual Greenshell™ mussels (Perna canaliculus) in a selective breeding programme. *Aquaculture,* 274**,** 41-48.
MASTRETTA-YANES, A., ARRIGO, N., ALVAREZ, N., JORGENSEN, T. H., PIÑERO, D. & EMERSON, B. C. 2015. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources,* 15**,** 28-41.
MCGOLDRICK, D. J. & HEDGECOCK, D. 1997. Fixation, Segregation and Linkage of Allozyme Loci in Inbred Families of the Pacific Oyster Crassostrea gigas (Thunberg): Implications for the Causes of Inbreeding Depression. *Genetics,* 146**,** 321-334.
NAKAMURA, K., OSHIMA, T., MORIMOTO, T., IKEDA, S., YOSHIKAWA, H., SHIWA, Y., ISHIKAWA, S., LINAK, M. C., HIRAI, A. & TAKAHASHI, H. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research,* 39**,** e90.
PARDO-MANUEL DE VILLENA, F. & SAPIENZA, C. 2001. Nonrandom segregation during meiosis: the unfairness of females. *Mammalian Genome,* 12**,** 331-339.
PELTOMAKI, P. 2001. Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Human Molecular Genetics,* 10**,** 735-40.
PLOUGH, L. V. 2012. Environmental stress increases selection against and dominance of deleterious mutations in inbred families of the Pacific oyster Crassostrea gigas. *Molecular Ecology,* 21**,** 3974-3987.
PLOUGH, L. V. 2016. Genetic load in marine animals: a review. *Current Zoology,* 62**,** 567-579.

PLOUGH, L. V. & HEDGECOCK, D. 2011. Quantitative Trait Locus Analysis of Stage-Specific Inbreeding Depression in the Pacific Oyster Crassostrea gigas. *Genetics,* 189**,** 1473-1486.

PLOUGH, L. V., SHIN, G. & HEDGECOCK, D. 2016. Genetic inviability is a major driver of type III survivorship in experimental families of a highly fecund marine bivalve. *Molecular Ecology,* 25**,** 895-910.

QUAIL, M. A., SMITH, M., COUPLAND, P., OTTO, T. D., HARRIS, S. R., CONNOR, T. R., BERTONI, A., SWERDLOW, H. P. & GU, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics,* 13**,** 341.

RAGG, N. L. C., KING, N., WATTS, E. & MORRISH, J. 2010. Optimising the delivery of the key dietary diatom Chaetoceros calcitrans to intensively cultured Greenshell™ mussel larvae, Perna canaliculus. *Aquaculture,* 306**,** 270-280.

REECE, K. S., RIBEIRO, W. L., GAFFNEY, P. M., CARNEGIE, R. B. & ALLEN, J. S. K. 2004. Microsatellite Marker Development and Analysis in the Eastern Oyster (Crassostrea virginica): Confirmation of Null Alleles and Non-Mendelian Segregation Ratios. *Journal of Heredity,* 95**,** 346-352.

RICHARDS, P. M., LIU, M. M., LOWE, N., DAVEY, J. W., BLAXTER, M. L. & DAVISON, A. 2013. RAD-Seq derived markers flank the shell colour and banding loci of the Cepaea nemoralis supergene. *Molecular Ecology,* 22**,** 3077-3089.

ROMIGUIER, J., GAYRAL, P., BALLENGHIEN, M., BERNARD, A., CAHAIS, V. & CHENUIL, A. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature.,* 515, 261-263.

RUMRILL, S. S. 1990. Natural mortality of marine invertebrate larvae. *Ophelia,* 32**,** 163-198.

THIRIOT-QUIEVREUX, C. 2002. Review of the literature on bivalve cytogenetics in the last ten years. *Cahiers de Biologie Marine,* 43**,** 17-26.

THORSON, G. 1950. Reproductive and larval ecology of marine bottom invertebrates. *Biological Reviews,* 25**,** 1-45.

THORSON, G. 1966. Some factors influencing the recruitment and establishment of marine benthic communities. *Netherlands Journal of Sea Research,* 3**,** 267-293.

WIENS, D., CALVIN, C. L., WILSON, C. A., DAVERN, C. I., FRANK, D. & SEAVEY, S. R. 1987. Reproductive success, spontaneous embryo abortion, and genetic load in flowering plants. *Oecologia,* 71**,** 501-509.

WILLIAMS, G. C. 1975. Sex and Evolution. *Princeton University Press, Princeton, NJ*.

WYGODA, J. A., YANG, Y., BYRNE, M. & WRAY, G. A. 2014. Transcriptomic analysis of the highly derived radial body plan of a sea urchin. *Genome Biol Evol,* 6**,** 964-73.

ZHANG, G., FANG, X., GUO, X., LI, L., LUO, R., XU, F., YANG, P., ZHANG, L., WANG, X., QI, H., XIONG, Z., QUE, H., XIE, Y., HOLLAND, P. W. H., PAPS, J., ZHU, Y., WU, F., CHEN, Y., WANG, J., PENG, C., MENG, J., YANG, L., LIU, J., WEN, B., ZHANG, N., HUANG, Z., ZHU, Q., FENG, Y., MOUNT, A., HEDGECOCK, D., XU, Z., LIU, Y., DOMAZET-LOSO, T., DU, Y., SUN, X., ZHANG, S., LIU, B., CHENG, P., JIANG, X., LI, J., FAN, D., WANG, W., FU, W., WANG, T., WANG, B., ZHANG, J., PENG, Z., LI, Y., LI, N., WANG, J., CHEN, M., HE, Y., TAN, F., SONG, X., ZHENG, Q., HUANG, R., YANG, H., DU, X., CHEN, L., YANG, M., GAFFNEY, P. M., WANG, S., LUO, L., SHE, Z., MING, Y., HUANG, W., ZHANG, S., HUANG, B., ZHANG, Y., QU, T., NI, P., MIAO, G., WANG, J., WANG, Q., STEINBERG, C. E. W., WANG, H., LI, N., QIAN, L., ZHANG, G., LI, Y., YANG, H., LIU, X., WANG, J., YIN, Y. & WANG, J. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature,* 490**,** 49-54.

CHAPTER 7

General discussion

7.1 Introduction

Bivalve genetic research is characterised by the frequent presence of genetic markers showing distortions from expected genotype ratios. A widely held hypothesis to explain segregation distortions (SD) is recessive/dominant deleterious alleles in linkage with the interrogated markers, which are expressed at early larval development (Plough, 2016, Plough and Hedgecock, 2011, Plough et al., 2016, Bierne et al., 1998, Launey and Hedgecock, 2001). It has been suggested that the origin of these deleterious alleles is a high rate of mutations that is fuelled by a high gametic output. The main supporting evidence for the hypothesis that SDs are caused by mutations is the substantial genotype dependant mortalities of bivalve larvae. At early larval stages, markers show no distortions from the expected genotype proportions (Bierne et al., 1998, Launey and Hedgecock, 2001), suggesting that the appearance of SD later in development is the by-product of genotype-dependant selection. As most of this genotype-dependant selection occurs at the larval stage, the analysis of this life cycle is critical for unraveling fundamental aspects of SD patterns in bivalves. In this thesis methods for high-throughput genotyping were used to provide insights into the origin of SD in bivalves. Patterns of SD were characterized at an unprecedented high resolution in the Pacific oyster *Crassostrea gigas* genome with the aid of a 55K SNP chip (Gutierrez et al., 2017) and the RAD-Seq technique (Baird et al., 2008, Miller et al., 2007). While in Greenshell^{TM} mussel *Perna canaliculus* families, sequencing technologies were used to uncover potential sources of *de novo* mutations. Finally, to get an estimate of the genotyping errors that may result from the analysis of a species with a complex genome, a detailed comparison of two commonly-used high-throughput genotyping technologies (SNP array vs. RAD-Seq method) is performed. In the following section, the main findings of the research are discussed in the context of existing literature.

7.2 Temporal analysis of SD

The literature indicates that in marine bivalves SD appears during larval development as a result of natural selection acting on linked deleterious mutations (Bierne et al., 1998, Launey and Hedgecock, 2001, Plough and Hedgecock, 2011). As embryonic and larval development progresses, genes are sequentially turned on for the first time. Hence, if a mutation disrupts the function of a relevant gene that is normally expressed at a certain

point in development, the viability of the individual may be comprised. Therefore, an indirect estimation of the presence of detrimental mutations may be achieved by measuring its consequence, that is, genotype-dependant mortalities. If a mutation is exerting a negative effect, then significant changes in allele frequencies should be observed when comparing the genetic pools of two consecutive larval stages.

The temporal analysis performed in this thesis revealed that the strongest selection pressure occurred between the (virtual) time-zero embryos and the first larval stage sampled, that is, the trochophore stage (~20hpf). Contrary to our finding, three studies agree in that at the trochophore stage no significant departures from Mendelian expectations are observed (i.e., the absence of SD) (Bierne et al., 1998, Launey and Hedgecock, 2001, Plough and Hedgecock, 2011). In these studies, SD mostly appears later in larval development, near metamorphosis, a critical transition at which the animal changes from a free-swimming planktonic stage to a juvenile with a benthic life. Some reasons for this discrepancy may be that previous studies genotyped individual larvae with a maximum of 80 microsatellites, whereas we sequenced pools of Greenshell™ mussel larvae and genotyped >1,300 RAD-loci (each 100bp in length) across families. The greatest advantage of pooling versus single individual analysis is that more individuals can be simultaneously genotyped (around 130 larvae by pool in this study). However, since by sequencing in pools we lose the link between an individual and its genotype, the analysis of SD is based on allelic frequencies in pools of larvae instead of the evaluation of genotype proportions in a group of individuals. To simplify the analysis, we created a 'virtual' pool that represented the allele ratios that should have been expected under normal Mendelian inheritance. This 'virtual' pool was generated by merging the RAD-Seq data of the parental gametes, after normalization. The fact that we used an artificial instead of a biological sample as the zero time point complicates the interpretation of the high distortions detected at the window between fertilization and the first larval stage sampled at ~20 hpf. For instance, it may be that mutations cause gamete inviability. This differential viability of gametes could be caused, for example, by deleterious mutations affecting gamete functionality such as motility in spermatozoa (Takasaki et al., 2014). Therefore, our assumption that all gametes have the potential of becoming embryos is invalid. Another possibility that would have led to the incorrect base-line allelic ratios in the 'virtual' pool is the unequal representation of parental alleles (e.g., caused by meiotic drive). However, no preferential transmission of any parental allele was detected in the Greenshell™ mussel

broodstock. Therefore, the presence of transmission ratio distortion was discarded. Ideally, the first time point of the temporal analysis should have been represented by pools of embryos sampled immediately after fertilization. However, this approach is impracticable, as embryos have the same size than unfertilized eggs. Hence, they cannot readily be separated from the gametes, limiting the possibility of obtaining a reliable estimate of allele ratios in pools of embryos.

Under the assumption that the allele frequencies of the 'virtual' pool are a close representation of the true allele ratios of embryos, we observe that most SD occur at the early (embryonic) stage. These results suggest that a substantial amount of mussel embryos carry genetic combinations that are inviable, similar to what has been observed in long-lived plants (Sorensen, 1969). Further studies could perform whole transcriptome sequencing on different larval stages to evaluate the stage-specific expression of deleterious alleles and their effect on fitness. In a broader context, it would be interesting to evaluate whether molluscs that have different reproduction strategies (e.g., brooding gastropods) share similar mutation rates than bivalves. This inter-class comparison would provide direct evidence of the evolutionary importance of mutation rates as an adaptive mechanism to cope with the fluctuating, stressful marine environments.

7.3 Genome-wide patterns of SD

SD is suggested to be a driving force of evolution (Lyttle, 1991). Markers showing distorted segregation ratios have been reported since the beginning of bivalve genetic research (Beaumont, 1991, Borsa et al., 1991, David et al., 1997). Several mechanisms have been proposed to explain SD, including gamete selection, preferential transmission of alleles or post-zygotic selection. In marine bivalves, the transmission of recessive/dominant (lethal and sub-lethal) mutations has been suggested as a possible cause of SD (Plough and Hedgecock, 2011, Plough et al., 2016, Launey and Hedgecock, 2001). SD patterns in bivalves have been previously studied with a moderate marker density (e.g., 99 markers in Plough et al. (2016)). The use of a higher marker density would enable the study of genetic phenomena such as SD at a higher resolution, providing insight into the origin of these unusual patterns.

In this thesis, genome-wide patterns of SD were studied in three Pacific oyster families by genotyping individuals with a higher density of markers. Two genotyping technologies were applied: a SNP array (~55k) and the RAD-Seq method. The percentage of highly distorted markers ranged between ~14% and 34% across the three families. To compare the SD patterns across families, SNPs from the array were placed on scaffolds of the Pacific oyster genome assembly (Zhang et al., 2012).

The majority of distorted markers clustered in specific regions of the chromosome, as observed in previous studies (Hedgecock et al., 2015). The physical location of regions showing SD (SDRs) was highly consistent between two of the three families analyzed. These families were (to the best of our knowledge) not closely related. On the other hand, the two families that shared the same dam show a low consistency of SDRs. Between these related half-sibling families, SDRs were commonly present in only one of the eight LGs analysed. Thus, the level of genetic difference between parents may be one of the causes of the difference in the number of SDRs detected among families. Although a high consistency between SDRs has not been detected previously, which was likely enabled in this study by the higher marker density utilized, SDs have been shown to cluster in specific LGs (Jones et al., 2013, Li and Guo, 2004, Hedgecock et al., 2015, McGoldrick and Hedgecock, 1997). In bivalve linkage studies, SD is mostly attributed to a deficiency of homozygote genotypes in the progeny (Li and Guo, 2004), evidence that suggests selection against linked deleterious recessive alleles. In this study, a tendency towards homozygote deficiencies is only observed in crosses in which the dam is heterozygous and the male homozygous. Notably, most SDRs tend to be biased towards the female parent, which is consistent with SD data obtained from a different oyster species (Jones et al., 2013). The observation that segregation ratios are skewed towards the female parent is consistent with the results from the evaluation of paternal and maternal allelic effects. We find that maternal alleles are responsible for the largest proportion of SDs in haplotype markers. Additional studies on multiple families created from parents with diverse genetic backgrounds (related vs. unrelated) would reveal the influence of female parents on SD, and whether it originates at the gamete or post-zygotic stage.

7. 4 Potential sources of *de novo* mutations in bivalves

The hypothesis that SD is caused by recessive/dominant mutations requires the demonstration of a high mutation rate. No direct incidence of a high mutation rate in a bivalve species has been provided, although preliminary data support this notion (Peñaloza, 2013). Nevertheless, it is worth mentioning that family-based studies have consistently detected the presence of 'contaminant' individuals at low frequencies (McGoldrick and Hedgecock, 1997), which may or may not have been a mutant.

The findings from this thesis suggest that *de novo* mutations are common in bivalves. Moreover, we detected that they occurred both in the germline and post-zygotically in the larvae. De novo mutations of post-zygotic origin were detected in all larval stages analyzed, from the trochophore (~20hpf) to the day 12 veliger larvae. The level of putative *de novo* single-nucleotide changes ranged from 207 to ~6,000. The frequency of these mutations was consistently low across the pool of larval offspring, ranging from 2% to 5% across families. Undoubtedly these values are extreme and would require confirmation with an orthogonal sequencing technology. Of particular concern was the result obtained from the experiment in which *de novo* mutations were tracked during development. The purpose of tracking the fate of the putative *de novo* mutations was to establish their nature – deleterious, sub-lethal, neutral or potential error. The behavior of post-zygotic mutations was odd, as they tended to appear only once in development, being completely absent in all previous and posterior larval stages. This observation contradicts the expectation of a neutral mutation (types of mutations that should be comparatively more abundant), which should remain present during development once it appears. A relevant factor that may be explaining the ephemeral nature of these mutations is the fact that they occur at low frequency. A study suggests that to confidently detect alleles at a frequency <5%, at least 250x coverage is required (Stead et al., 2013). The average read-depth per locus of our experiment was 90x per sample. Hence, repeating the experiment with more input DNA, sequenced at a higher coverage would provide unbiased evidence of the rate of post-zygotic mutations. Notably, if we conservatively establish that true post-zygotic *de novo* mutations are those that are present in at least three consecutive larval stages, then post-zygotic mutations would range from 254 to 871 across families. Given that we estimate that we are sequencing approximately 0.1% of the mussel genome, these conservative values are still high as an overall estimate.

An additional source of mutations in the Greenshell™ mussel was the germline. De novo mutations of germline origin were detected in both male and female gametes. Contrary to the expectation that mutation rates are more frequent in males due to their higher number of cell divisions, females showed the highest rates of mutations. The average number of mutations across females was 222, eightfold higher than the number of events detected in males (~27). The frequency of these mutations in the gamete pool was also higher for females compared to males (4% vs. 2%, respectively). A fraction (~10%) of the mutations that were observed in the female gametes were also found in the first larval stage, indicating that they contribute to the genetic pool, at least in early stages of development. The females (dams) that had the greatest number of germline *de novo* mutations were also the ones that contributed the most to the larval pool. Considering the hypothesis that SD is caused by deleterious mutations, then the higher rates of novel alleles in the maternal line should bias SDs towards the female parent. This hypothesis is consistent with the genome-wide analysis of SDRs performed in this thesis, as near twice of the regions showing distortions in Pacific oyster families were female based. Altogether the evidence suggests that females may play an important role in determining patterns of SD.

In summary, bivalves show a high rate of putative *de novo* mutations, as predicted from their life-history traits (Plough, 2016). Because of the scale of our experiment (focused on the larval cycle) we were not able to accurately determine the contribution of these mutations to the population. Ideally, a single experiment should have been designed to identify mutations (in both gametes and larvae) and monitor their fate, not only at the larval stage but at least until the reproductive maturity of the organism, mainly to determine if the *de novo* mutations are transmitted to the next-generation. This attempt was made on two occasions but was unsuccessful because of unexpectedly high larval mortalities encountered in the hatchery. Further experiments addressing the fate of mutations are required to understand the adaptive value of the putative high mutation rate of bivalves. For instance, the answer to whether a high mutation rate has an evolutionary advantage may be addressed performing multigenerational studies in divergent environments. Of particular utility for the aquaculture industry would be to evaluate the short-term adaptation of bivalves to algal blooms (Bricelj et al., 2005) or ocean acidification (Thomsen et al., 2017), two major environmental challenges for future sustainable production.

7.5 Genomic technologies in polymorphic species

Genotyping accuracy is crucial for genetic analysis, although the tolerable error rate will depend on the biological question under study. The bivalve genome is highly polymorphic (Sauvage et al., 2007, Zhang et al., 2012) and therefore at risk of technical errors such as null alleles. Indeed, null alleles are common in bivalve genetic studies (Launey and Hedgecock, 2001). Moreover, initially, they were considered a main responsible factor for SD when deviations were due to a deficiency of heterozygote genotypes in population studies (Foltz, 1986, Hare et al., 1996). Due to the importance of accurate genotyping for downstream analyses, the assessment of the genotyping performance of current technologies in species with challegning genomes is required.

Three Pacific oyster families were genotyped with two commonly used genotyping platforms – a SNP array (Gutierrez et al., 2017) and the RAD-Seq method (Miller et al., 2007, Baird et al., 2008). The genotyping rate of the platforms was calculated based on the proportion of markers showing Mendelian errors. Mendelian errors are defined when the genotype of an individual is inconsistent with the parental genotypes based on Mendelian rules of inheritance. The number or Mendelian errors detected for both platforms was high. When the families were genotyped with the SNP array, 8% of the markers showed Mendelian errors. Likewise, Qi et al. (2017) found high Mendelian error rates in the same species when families were genotyped with a 190k SNP array. On the other hand, when the families were genotyped with the RAD-Seq method, 24% of the haplotype markers showed a Mendelian error. No mention regarding a high Mendelian error rates was made in a study that utilized the same technology in another bivalve species (Li and He, 2014). Interestingly, when we estimated the genotyping error of the RAD-Seq method using technical replicates (i.e., same DNA sample processed twice), the genotyping error rate (measured as the concordant genotypes between sample replicates) was lower and ranged from 0.4% to 3.6%. Therefore, it appears that error rates are inflated when they are estimated from pedigree data (as Mendelian error rates). The most plausible explanation for these high levels of Mendelian error rates is null alleles. The high frequency of null alleles is likely caused by the high sequence and structural polymorphism of bivalves, which in turn is probably driven by high mutation rates. Therefore, null alleles may be inherited or occur spontaneously. A limitation of the present study was that the markers genotyped with both technologies (SNP array and RAD-Seq method) did not overlap. Hypothetically, if both

technologies genotyped the same locus, then it would have been possible to gain insight into the source of the error. For instance, if the same markers showed Mendelian errors in both platforms, then the error would have been likely to be caused by biological factors. To truly clarify the origin of Mendelian errors in bivalves, genomes of different species should be studied at greater detail, ideally by using technologies that allow the phasing of each parental haplotype. Indeed, the sequencing of bivalve genomes with technologies such as 10x (www.10xgenomics.com) may help build a comprehensive view of the mutational landscape present in this group of species. Moreover, 10x has successfully been used as a method to study large rearrangements in cancer genomes (Xia et al., 2018), holding promise for the study of the complex bivalve genomes. Of importance for selective breeding would be to determine whether the inter-individual difference in germline mutation rate – which was more prominent in females – is heritable and can therefore be selected on as a production trait. In the same context, it would be interesting to explore the connection between the low levels of methylation found in marine bivalves with stochastic variation in transcriptional products (Gavery and Roberts, 2014), as observed in the honeybee *Apis mellifera* (Li-Byarlay et al., 2013). This avenue of research would inform on the importance of phenotypic plasticity for bivalve species and, indirectly, the extent to which selective breeding is feasible.

In summary, technical errors associated with genotyping performance are not comparatively high (based on the results from the technical replicates). This however does not imply that genotypes are reflecting the true diploid state. The high incidence of Mendelian errors, particularly observed in the RAD-Seq data, suggests (to some extent) the decoupling of segregation ratios from Mendelian expectation. One possibility would be that null alleles are segregating at a high frequency in the data. Attention must be paid in population genetic studies that utilize RAD-Seq for marker discovery and genotyping, especially if an excess of homozygote genotypes is encountered, as they may be driven by null alleles instead of reflecting population processes.

7.6 Conclusion

Next-generation sequencing technologies were used to uncover fundamental aspects of the longstanding phenomenon of SD in bivalve species. Moreover, for the first time, a genome-wide genotyping by sequencing method (RAD-Seq) was used to unravel the genetic composition of larval stages of a species with indirect development. High levels of null alleles were detected in bivalves, which are reflected by the unusual bimodal distribution of histograms of depth of coverage per locus. These null alleles are a likely causing the high frequency of Mendelian errors detected the RAD-Seq and SNP array data. The high density of genetic markers allowed us to study the genome-wide patterns of SD at a high resolution. Distorted markers tended to cluster in specific linkage groups. Consistency between SDRs was found between two unrelated families, suggesting SDRs may encompass viability-related genes. No clear tendency was observed in SDRs, as distortions were caused both by homozygote excess or deficiency. A feature of bivalves that may be contributing to the generation of these SDs is a high mutation rate. Direct evidence of a high mutation rate is provided for the first time, and the indicative data suggest they occur at a rate far in excess of what has been shown for other species. Two putative sources of *de novo* mutations were identified: the germline and posy-zygotic events. Because of the preliminary nature of this finding, further research is warranted to confirm this high rate of *de novo* mutations.

7.3 References

BAIRD, N. A., ETTER, P. D., ATWOOD, T. S., CURREY, M. C., SHIVER, A. L. & LEWIS, Z. A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One,* 3,e3376.

BEAUMONT, A. R. 1991. Genetic studies of laboratory reared mussels, Mytilus edulis: heterozygote deficiencies, heterozygosity and growth. *Biological Journal of the Linnean Society,* 44**,** 273-285.

BIERNE, N., LAUNEY, S., NACIRI-GRAVEN, Y. & BONHOMME, F. 1998. Early Effect of Inbreeding as Revealed by Microsatellite Analyses on Ostrea edulis Larvae. *Genetics,* 148**,** 1893-1906.

BORSA, P., ZAINURI, M. & DELAY, B. 1991. Heterozygote deficiency and population structure in the bivalve Ruditapes decussatus. *Heredity,* 66**,** 1-8.

BRICELJ, V., CONNELL, L., KONOKI, K., MACQUARRIE, S., SCHEUER, T., CATTERALL, W. & L TRAINER, V. 2005. Sodium channel mutation leading to saxitoxin resistance in clams increases risk of PSP. *Nature*, 763-767.

DAVID, P., PERDIEU, M. A., PERNOT, A. F. & JARNE, P. 1997. FINE-GRAINED SPATIAL AND TEMPORAL POPULATION GENETIC STRUCTURE IN THE MARINE BIVALVE SPISULA OVALIS. *Evolution,* 51**,** 1318-1322.

FOLTZ, D. W. 1986. NULL ALLELES AS A POSSIBLE CAUSE OF HETEROZYGOTE DEFICIENCIES IN THE OYSTER CRASSOSTREA VIRGINICA AND OTHER BIVALVES. *Evolution,* 40**,** 869-870.

GAVERY, M. R. & ROBERTS, S. B. 2014. A context dependent role for DNA methylation in bivalves. *Briefings in Functional Genomics,* 13**,** 217-222.

GUTIERREZ, A. P., TURNER, F., GHARBI, K., TALBOT, R., LOWE, N. R., PEÑALOZA, C., MCCULLOUGH, M., PRODÖHL, P. A., BEAN, T. P. & HOUSTON, R. D. 2017. Development of a Medium Density Combined-Species SNP Array for Pacific and European Oysters (Crassostrea gigas and Ostrea edulis). *G3: Genes|Genomes|Genetics*, 7, 2209-2218

HARE, M. P., KARL, S. A. & AVISE, J. C. 1996. Anonymous nuclear DNA markers in the American oyster and their implications for the heterozygote deficiency phenomenon in marine bivalves. *Molecular Biology and Evolution,* 13**,** 334-345.

HEDGECOCK, D., SHIN, G., GRACEY, A. Y., DEN BERG, D. V. & SAMANTA, M. P. 2015. Second-Generation Linkage Maps for the Pacific Oyster Crassostrea gigas Reveal Errors in Assembly of Genome Scaffolds. *G3: Genes|Genomes|Genetics,* 5**,** 2007-2019.

JONES, D. B., JERRY, D. R., KHATKAR, M. S., RAADSMA, H. W. & ZENGER, K. R. 2013. A high-density SNP genetic linkage map for the silver-lipped pearl oyster, Pinctada maxima: a valuable resource for gene localisation and marker-assisted selection. *BMC Genomics,* 14**,** 810.

LAUNEY, S. & HEDGECOCK, D. 2001. High genetic load in the Pacific oyster Crassostrea gigas. *Genetics,* 159**,** 255-265.

LI, L. & GUO, X. 2004. AFLP-Based Genetic Linkage Maps of the Pacific Oyster Crassostrea gigas Thunberg. *Marine Biotechnology,* 6**,** 26-36.

LI, Y. & HE, M. 2014. Genetic Mapping and QTL Analysis of Growth-Related Traits in Pinctada fucata Using Restriction-Site Associated DNA Sequencing. *PLOS ONE,* 9**,** e111707.

LI-BYARLAY, H., LI, Y., STROUD, H., FENG, S., NEWMAN, T. C., KANEDA, M., HOU, K. K., WORLEY, K. C., ELSIK, C. G., WICKLINE, S. A., JACOBSEN, S. E., MA, J. & ROBINSON, G. E. 2013. RNA interference knockdown of DNA methyl-transferase 3 affects gene

alternative splicing in the honey bee. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 12750-12755.

LYTTLE, T. W. 1991. Segregation Distorters. *Annual Review of Genetics,* 25**,** 511-581.

MCGOLDRICK, D. J. & HEDGECOCK, D. 1997. Fixation, Segregation and Linkage of Allozyme Loci in Inbred Families of the Pacific Oyster Crassostrea gigas (Thunberg): Implications for the Causes of Inbreeding Depression. *Genetics,* 146**,** 321-334.

MILLER, M. R., DUNHAM, J. P., AMORES, A., CRESKO, W. A. & JOHNSON, E. A. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research,* 17**,** 240-248.

PEÑALOZA, C. 2013. *Development and application of genomic tools to the genetic improvement of Atlantic salmon and Chilean mussels.* University of Edinburgh.

PLOUGH, L. V. 2016. Genetic load in marine animals: a review. *Current Zoology,* 62**,** 567-579.

PLOUGH, L. V. & HEDGECOCK, D. 2011. Quantitative Trait Locus Analysis of Stage-Specific Inbreeding Depression in the Pacific Oyster Crassostrea gigas. *Genetics,* 189**,** 1473-1486.

PLOUGH, L. V., SHIN, G. & HEDGECOCK, D. 2016. Genetic inviability is a major driver of type III survivorship in experimental families of a highly fecund marine bivalve. *Molecular Ecology,* 25**,** 895-910.

QI, H., SONG, K., LI, C., WANG, W., LI, B., LI, L. & ZHANG, G. 2017. Construction and evaluation of a high-density SNP array for the Pacific oyster (Crassostrea gigas). *PLOS ONE,* 12**,** e0174007.

SAUVAGE, C., BIERNE, N., LAPÈGUE, S. & BOUDRY, P. 2007. Single Nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster Crassostrea gigas. *Gene,* 406**,** 13-22.

SORENSEN, F. 1969. Embryonic Genetic Load in Coastal Douglas-Fir, Pseudotsuga menziesii var. Menziesii. *The American Naturalist,* 103**,** 389-398.

STEAD, L. F., SUTTON, K. M., TAYLOR, G. R., QUIRKE, P. & RABBITTS, P. 2013. Accurately Identifying Low-Allelic Fraction Variants in Single Samples with Next-Generation Sequencing: Applications in Tumor Subclone Resolution. *Human Mutation,* 34**,** 1432-1438.

TAKASAKI, N., TACHIBANA, K., OGASAWARA, S., MATSUZAKI, H., HAGIUDA, J., ISHIKAWA, H., MOCHIDA, K., INOUE, K., OGONUKI, N., OGURA, A., NOCE, T., ITO, C., TOSHIMORI, K. & NARIMATSU, H. 2014. A heterozygous mutation of GALNTL5 affects male infertility with impairment of sperm motility. *Proceedings of the National Academy of Sciences of the United States of America,* 111**,** 1120-1125.

THOMSEN, J., STAPP, L. S., HAYNERT, K., SCHADE, H., DANELLI, M., LANNIG, G., WEGNER, K. M. & MELZNER, F. 2017. Naturally acidified habitat selects for ocean acidification–tolerant mussels. *Science Advances,* 3.

XIA, L. C., BELL, J. M., WOOD-BOUWENS, C., CHEN, J. J., ZHANG, N. R. & JI, H. P. 2018. Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic Acids Research,* 46**,** e19-e19.

ZHANG, G., FANG, X., GUO, X., LI, L., LUO, R., XU, F., YANG, P., ZHANG, L., WANG, X., QI, H., XIONG, Z., QUE, H., XIE, Y., HOLLAND, P. W. H., PAPS, J., ZHU, Y., WU, F., CHEN, Y., WANG, J., PENG, C., MENG, J., YANG, L., LIU, J., WEN, B., ZHANG, N., HUANG, Z., ZHU, Q., FENG, Y., MOUNT, A., HEDGECOCK, D., XU, Z., LIU, Y., DOMAZET-LOSO, T., DU, Y., SUN, X., ZHANG, S., LIU, B., CHENG, P., JIANG, X., LI, J., FAN, D., WANG, W., FU, W., WANG, T., WANG, B., ZHANG, J., PENG, Z., LI, Y., LI, N., WANG, J., CHEN, M., HE, Y., TAN, F., SONG, X., ZHENG, Q., HUANG, R., YANG, H., DU, X., CHEN, L., YANG, M., GAFFNEY, P. M., WANG, S., LUO, L., SHE, Z., MING, Y., HUANG, W., ZHANG, S.,

HUANG, B., ZHANG, Y., QU, T., NI, P., MIAO, G., WANG, J., WANG, Q., STEINBERG, C. E. W., WANG, H., LI, N., QIAN, L., ZHANG, G., LI, Y., YANG, H., LIU, X., WANG, J., YIN, Y. & WANG, J. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature,* 490**,** 49-54.

Appendix

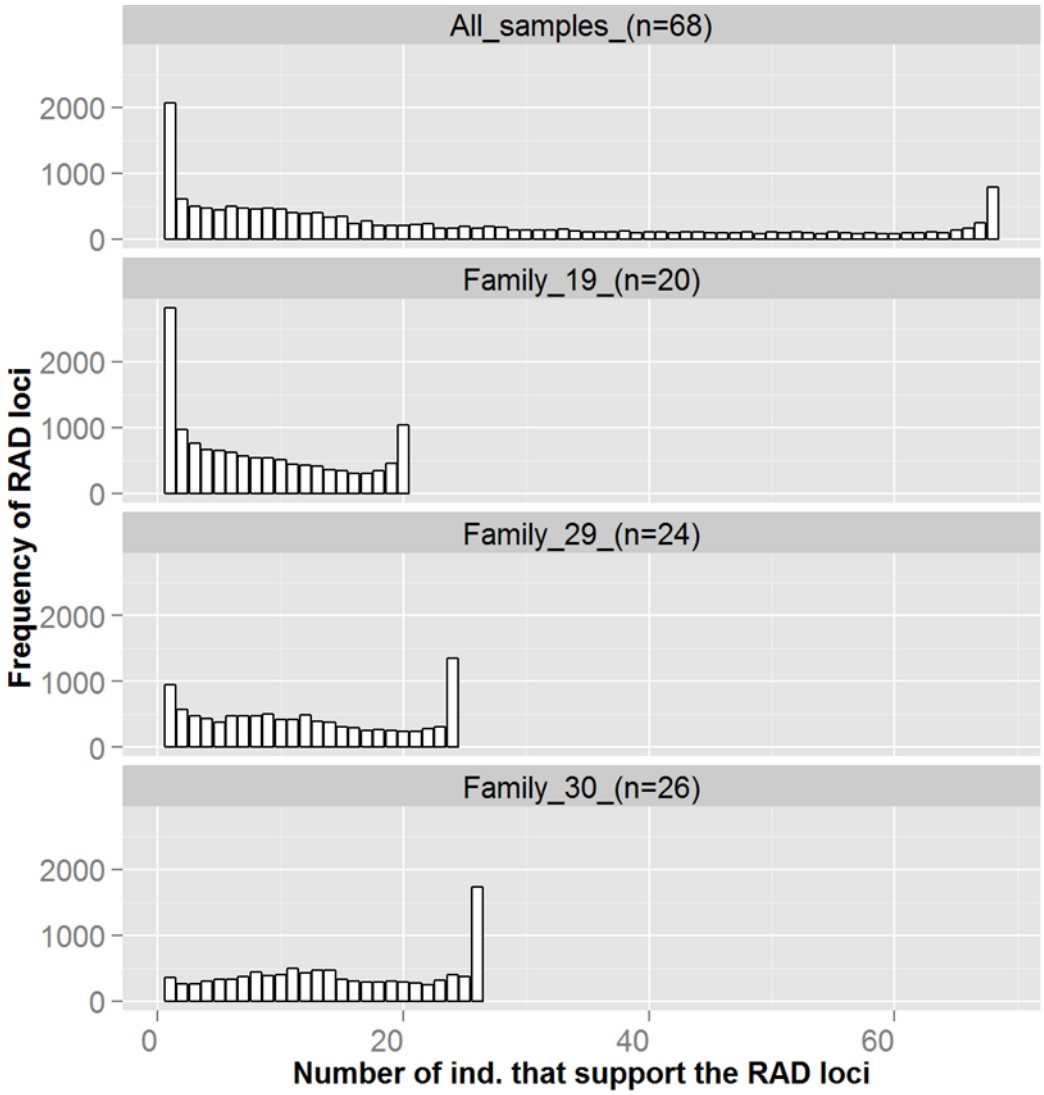


Figure 1. Reduced number of RAD-loci identified in a catalogue of loci created from Pacific oyster individuals across three families (top graph) compared to catalogues of loci created only from family members. In the y-axes the frequency of RAD-loci, and in the x-axes the number of individuals that align (support) a specific RAD-loci in a catalogue.

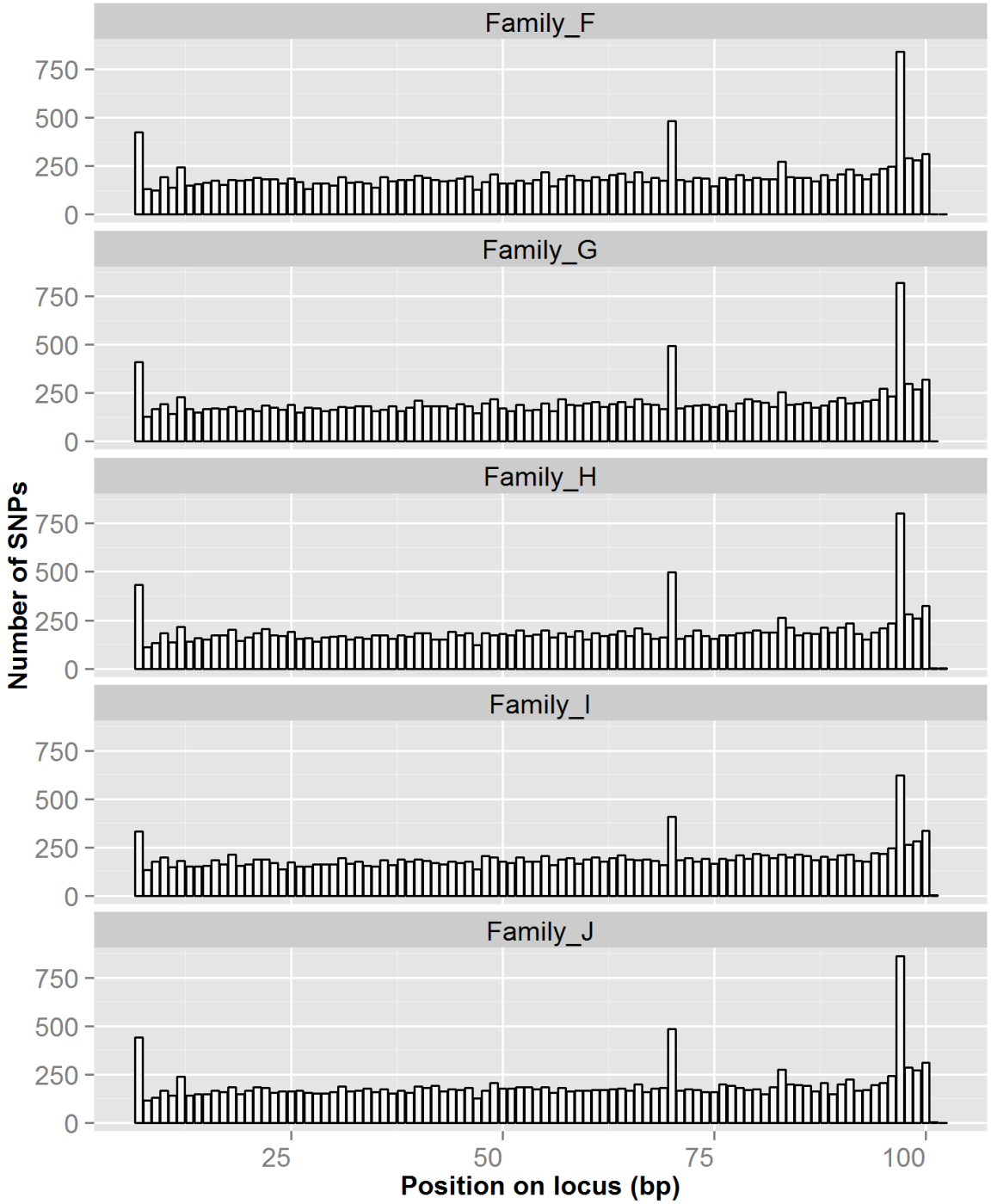
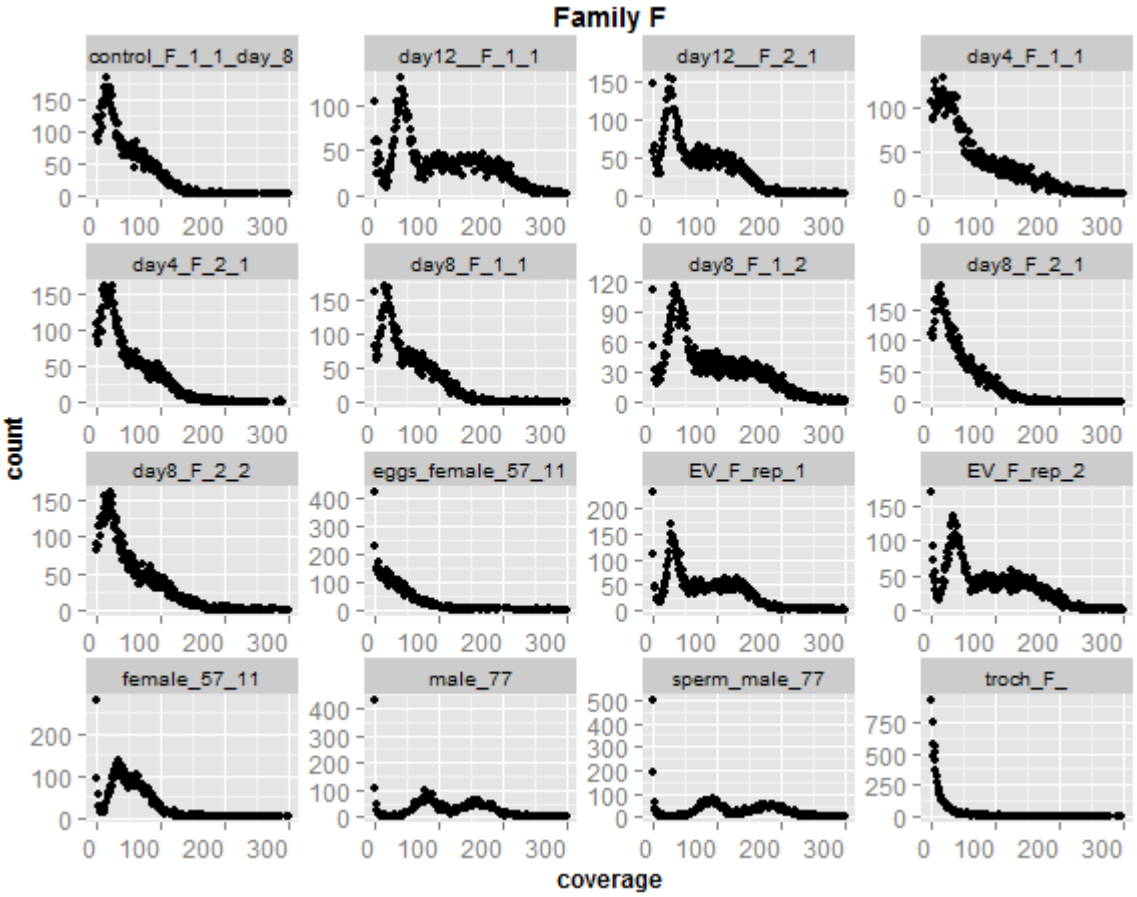
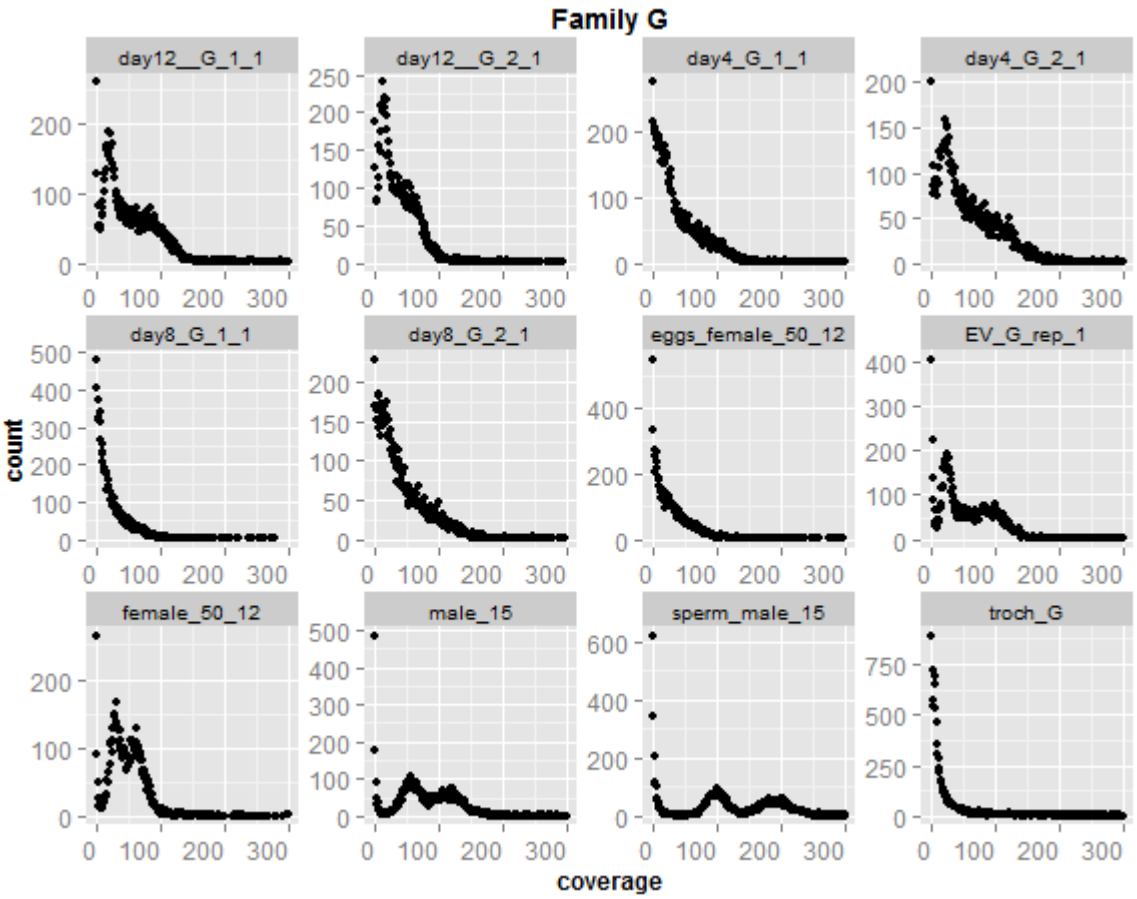
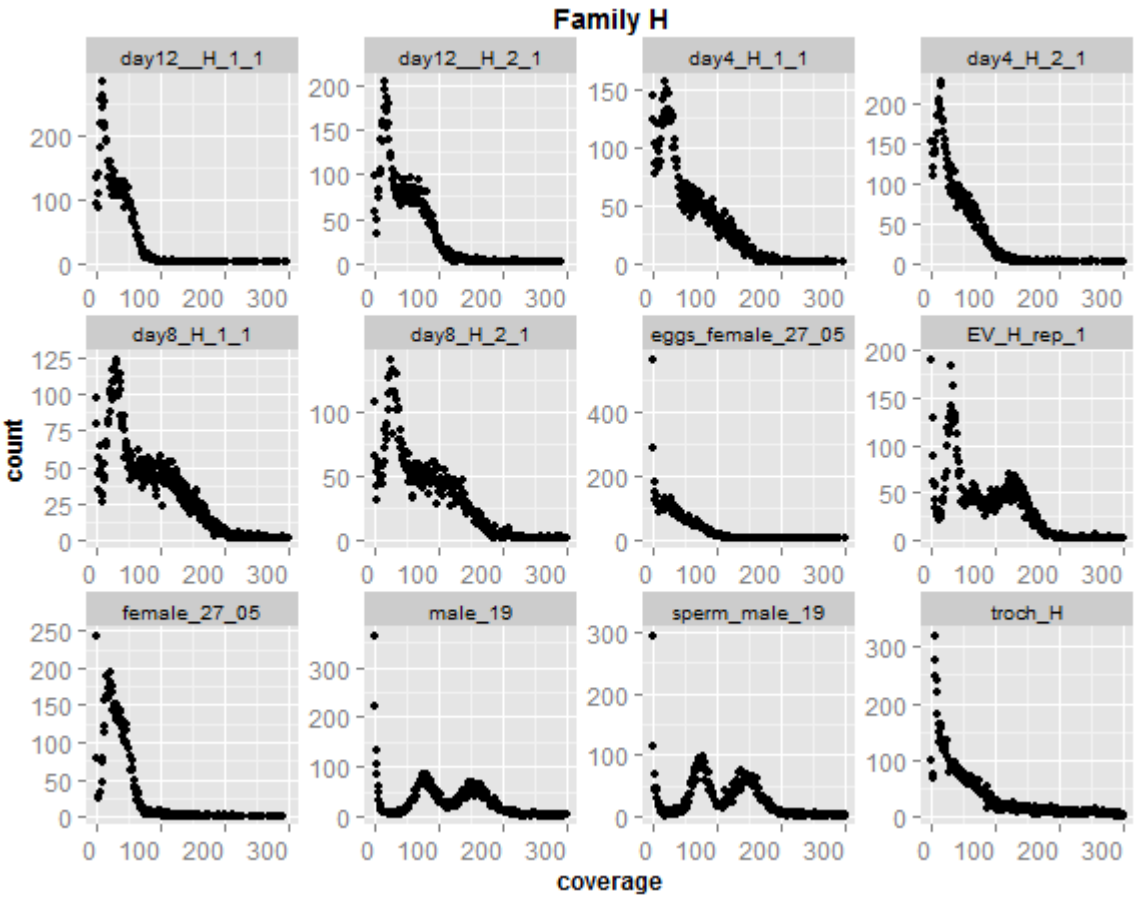
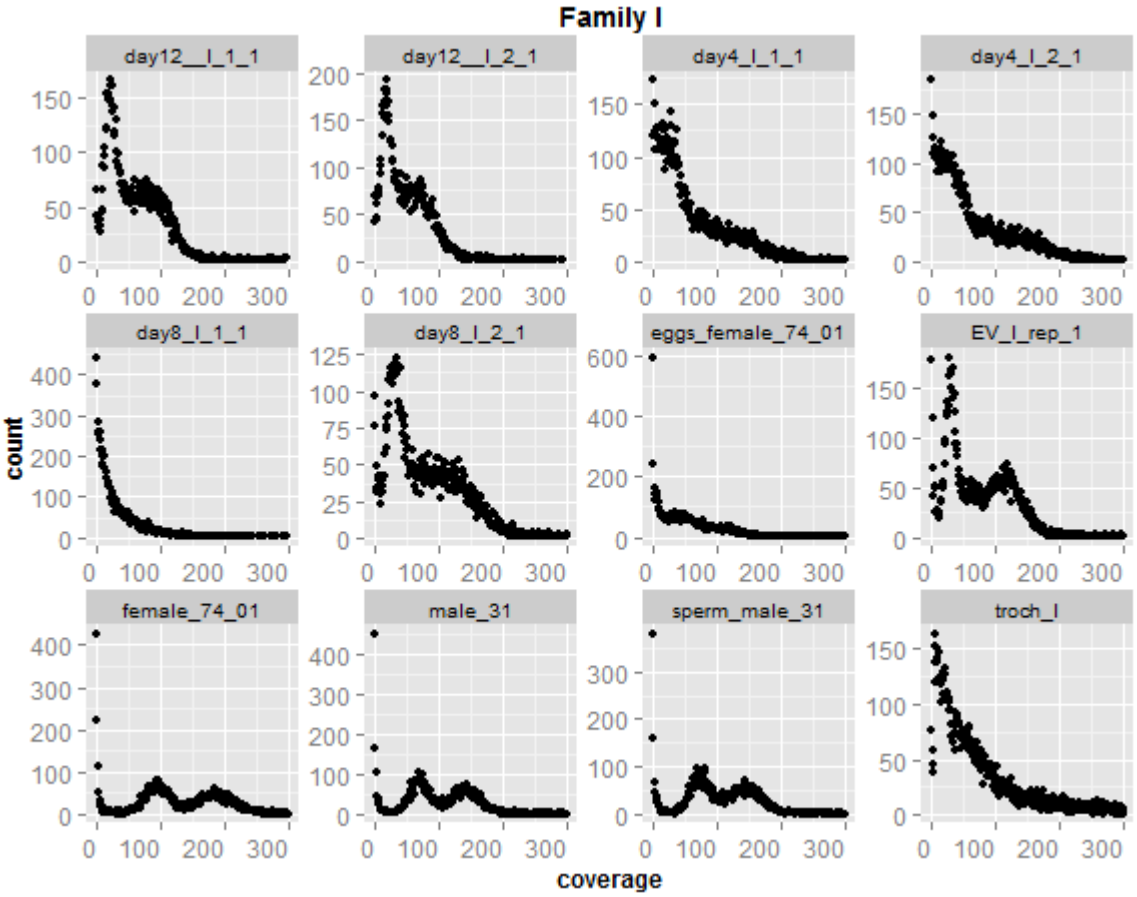


Figure 2. Unexpected increase of SNP counts at three nucleotide positions on the RAD-loci. Data are presented separately for each Greenshell™ mussel family.


Family F
control_F_1_1_day_8
day12__F_1_1
day12__F_2_1
day4_F_1_1
day4_F_2_1
day8_F_1_1
day8_F_1_2
day8_F_2_1
day8_F_2_2
eggs_female_57_11
EV_F_rep_1
EV_F_rep_2
female_57_11
male_77
sperm_male_77
troch_F_
count
coverage


Family G
day12__G_1_1
day12__G_2_1
day4_G_1_1
day4_G_2_1
day8_G_1_1
day8_G_2_1
eggs_female_50_12
EV_G_rep_1
female_50_12
male_15
sperm_male_15
troch_G
count
coverage


Family H
day12__H_1_1
day12__H_2_1
day4_H_1_1
day4_H_2_1
day8_H_1_1
day8_H_2_1
eggs_female_27_05
EV_H_rep_1
female_27_05
male_19
sperm_male_19
troch_H
count
coverage


Family I
day12__I_1_1
day12__I_2_1
day4_I_1_1
day4_I_2_1
day8_I_1_1
day8_I_2_1
eggs_female_74_01
EV_I_rep_1
female_74_01
male_31
sperm_male_31
troch_I
count
coverage

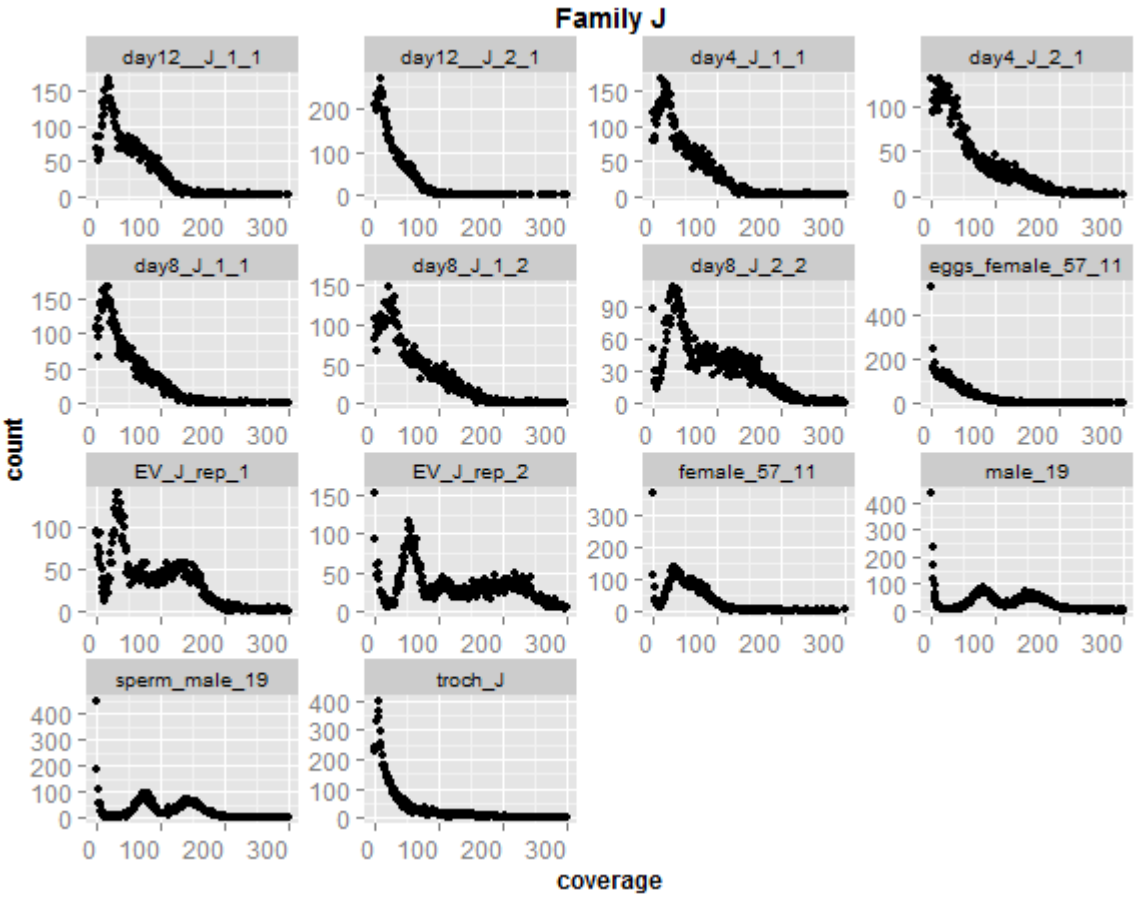


Figure 3. Frequency of locus read-depth across different Greenshell™ mussel samples and families. Read depth is the total number of reads from a locus. In the x-axes the number of RAD-loci, and in the y-axes the coverage per locus. A coverage range between 1 and 300 is shown for convenience.

Section 2

RAD sequencing of whole-genome amplified single mussel larvae

Introduction

An approach to characterize genome-wide patterns of inheritance in larval offspring is to genotype single larva at a high marker density. The technical limitation of this approach is that larvae are microscopic (<200um) and their DNA yield is insufficient for sequencing with a technique such as RAD-Seq. To circumvent this issue, the amount of starting material may be increased by whole genome amplification (WGA) techniques. To be useful for molecular analysis, a WGA technique must provide an unbiased representation of the genome, which has to be amplified with high fidelity.

Multiple displacement amplification (MDA) is a non-PCR based WGA technique used in the amplification of very low DNA quantities (Dean et al., 2002). In MDA, DNA is amplified isothermally (i.e., at a constant temperature) by using the high fidelity Φ29 DNA polymerase. MDA-based WGA is favoured over other PCR-based WGA techniques regarding bias amplification and reproducibility (Hosono et al., 2003). In this experiment, MDA was used to amplify the genomic DNA of single larvae, which was then used for the preparation of a RAD-Seq library to study marker inheritance patterns in blue mussel *Mytilus edulis* larval offspring.

Material and Methods

Pair-crosses

Twenty-four blue mussel *M. edulis* families were produced in 2014 at the Centre for Environment, Fisheries and Aquaculture Science (Cefas), Weymouth, the UK (see Material and Methods Chapter). Tissue samples were taken from the parents at the time of fertilisation and preserved in absolute ethanol. Samples from larval offspring were taken at 15 and 72 hour after fertilisation. The different larval stages were cryopreserved in 10%

DMSO and transferred to The Roslin Institute, UK, for DNA extraction, genotyping and analysis.

DNA extraction

Genomic DNA was extracted from the adult tissue using a CTAB-based extraction method (Richards et al., 2013). To extract DNA from 72 hour post fertilisation larval offspring, single larvae were isolated manually under a microscope using a 10ul micropipette. Each larva was washed 3 times in a Sodium Chloride 0.9% solution to remove debris that may contaminate the DNA extraction. The larva was transferred to an autoclaved 1.5ml Eppendorf tube and the presence of a single individual was confirmed by visual inspection under a microscope. DNA was extracted from the larval offspring by adding 60ul of Chelex 5% and 5ul of proteinase K (20mg/ml) to each tube and leaving the sample incubating overnight at 37°C. The following day extractions were centrifuged at 1,000 rpm and the supernatant transferred to an autoclaved 1.5ml Eppendorf tube. The DNA extraction was stabilized by adding an equal volume of Buffer EB (Qiagen).

WGA of larvae

After decontaminating all surfaces and equipment with 5% sodium hypochlorite, the genome of individual larvae was amplified using a MDA-based WGA kit (REPLI-g Single Cell Kit, Quiagen), following the manufacturer's instruction. Five microliters from each larval DNA extraction were used in the amplification reaction. A negative control was included to assess the possibility of contamination of the reagents. All WGA DNA reactions were run on an agarose gel to evaluate the quality of the amplification. To confirm that the WGA DNA was of mussel origin, a PCR was carried out to amplify a mitochondrial (COI) and a nuclear locus (Me15/Me16).

RAD-Seq library preparation

The genomic DNA from three blue mussel families was used for library preparation. Each family included the parents and ~40 larval offspring whose genome was previously amplified using MDA. RAD libraries were prepared following a protocol described in the Material and Methods Chapter of this thesis.

Additionally, as a means of assessing the technical variation introduced by MDA and the library preparation procedure, a set of controls were included:

Control 1: The DNA from and adult was diluted (through a serial dilution) to ~5ng in order to match the theoretical amount of DNA being amplified from a single larva. Subsequently, the diluted DNA sample was amplified using MDA. Both the (i) undiluted (adult) DNA and (ii) diluted DNA that underwent MDA, were included in the preparation of the RAD-Seq library.

Control 2: The genomic DNA from an individual larva was amplified using MDA in two independent occasions (technical replicate for MDA amplification).

Control 3: The gDNA from a single individual, which was previously amplified by MDA, was processed as two independent samples during the library preparation process (technical replicate of the library preparation procedure).

Preliminary *de novo* assembly and SNP calling

De novo assembly of blue mussel single-end reads was performed using Stacks (Catchen et al., 2013, Catchen et al., 2011). Within individuals, reads were assembled into RAD-loci with the *ustacks* module using a minimum stack depth of 10, and a maximum number of mismatches between stacks (or the two putative alleles of an individual) of 6. In addition, we allowed for the presence of indel variation using the '--gapped-alignment' command. A catalogue of representative loci across individuals was obtained by merging individually assembled loci using a maximum sequence mismatch of 8. The *population* module in Stacks was used to generate a dataset of SNP markers genotyped across individuals if the locus had a minimum of 8 reads.

Results and discussion

The number of paired-end reads obtained per sample ranged from 52,495 to 2,445,998 with an average of 938,409 (SD =606,312). After the quality control filtering and read alignment, we found that the number of RAD-loci genotyped across larval offspring was significantly lower than the number of loci genotyped in the parents (adults) of the mussel families (Figure 1A). In average the number of loci in the parents was of 10,615, whereas the average number of loci in the offspring was 2,366. This difference in the number of loci between parents and offspring was not associated with the number of reads per sample, as shown in Figure 1B (Pearson correlation coefficient = 0.13).

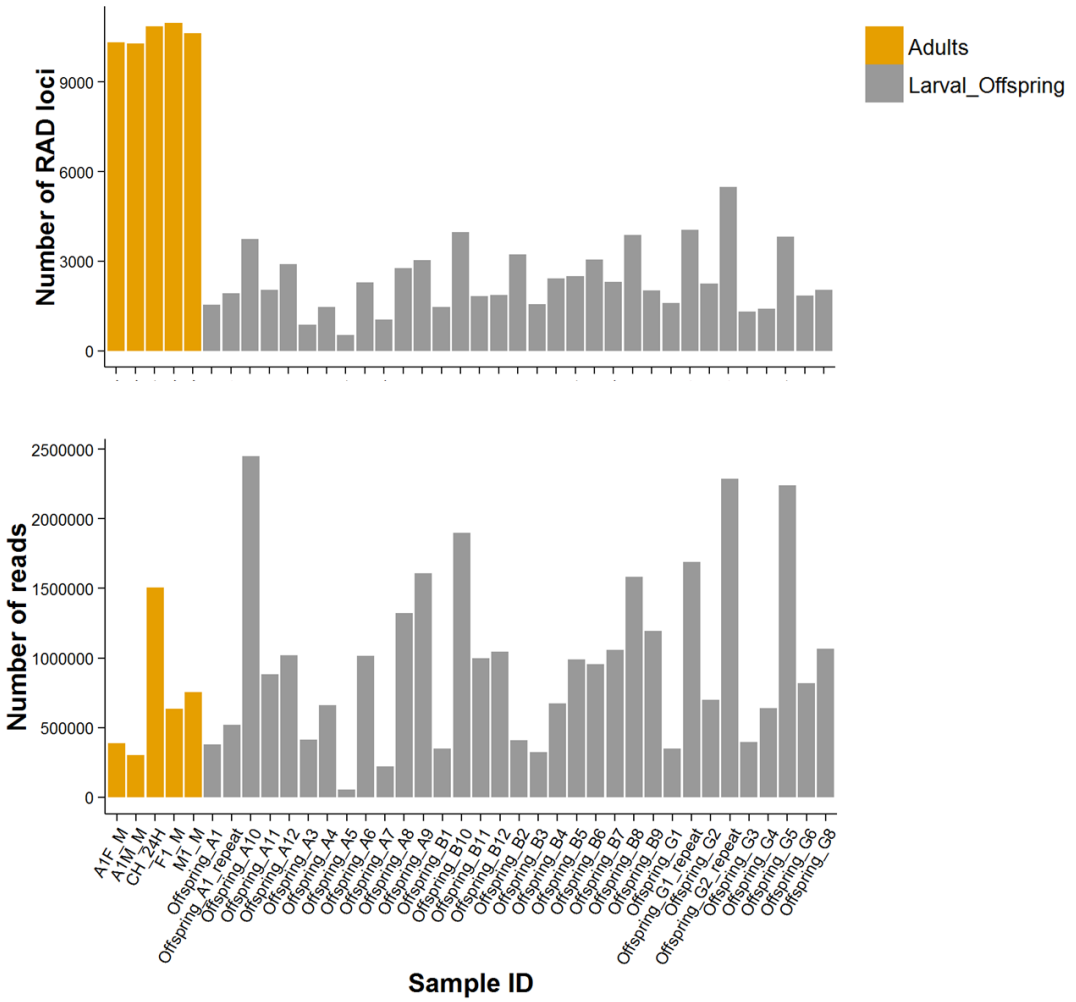


Figure 1. No relationship between the number of RAD-loci and the number of sequencing reads per sample was found.

The reason for the lower number of RAD-loci sequenced across the larval offspring was inferred from the analysis of the controls included in the RAD-Seq library. Considering the pair of technical replicates represented by an adult sample and its dilution that underwent MDA-based WGA (control 1), we observe a comparatively lower average coverage per locus for the diluted sample (Figure 2A). Additionally, the number of RAD-loci identified in the adult samples was higher than in the diluted replicate sample (8,817 vs. 4,008), as shown in the Venn diagram (Figure 2B). This suggests that the gDNA of the diluted adult mussel sample may have been amplified with a certain degree of bias, given not all loci appear to have been successfully sequenced in the diluted sample. The potential bias introduced by WGA becomes more evident in the technical replicates of the MDA reaction (i.e., same individual larva amplified twice by MDA) (control 2). We observe, apart from the overall low coverage per locus for both sample replicates (Figure 3A), a low number of shared RAD-loci (Figure 3B). This is evidence in favour of a biased whole genome amplification procedure, as RAD-loci appear to have been inconsistently amplified in both replicates of a pair, leading to a low number of common sequenced RAD-loci. The hypothesis that MDA may have amplified the larval genome in a biased manner is further confirmed by comparing the technical replicate for the library preparation procedure (control 3). Similar to the other controls, we observe a low coverage per locus, which at this stage appears to be typical of samples that are amplified by MDA prior to library construction. For this replicate pair, however, most of the identified RAD-loci within each sample were also common between samples, implying a low technical variation derived from the preparation of the RAD-Seq library. Altogether the evidence suggests that all the bias observed in the RAD-loci sampling occurred before library preparation. Overall, the RAD-Seq of WGA larval DNA suggests that the amplified larval genomic DNA was a biased representation of the original sequence. Consequently, future analysis of larval stages is performed by sequencing pools of larvae instead.

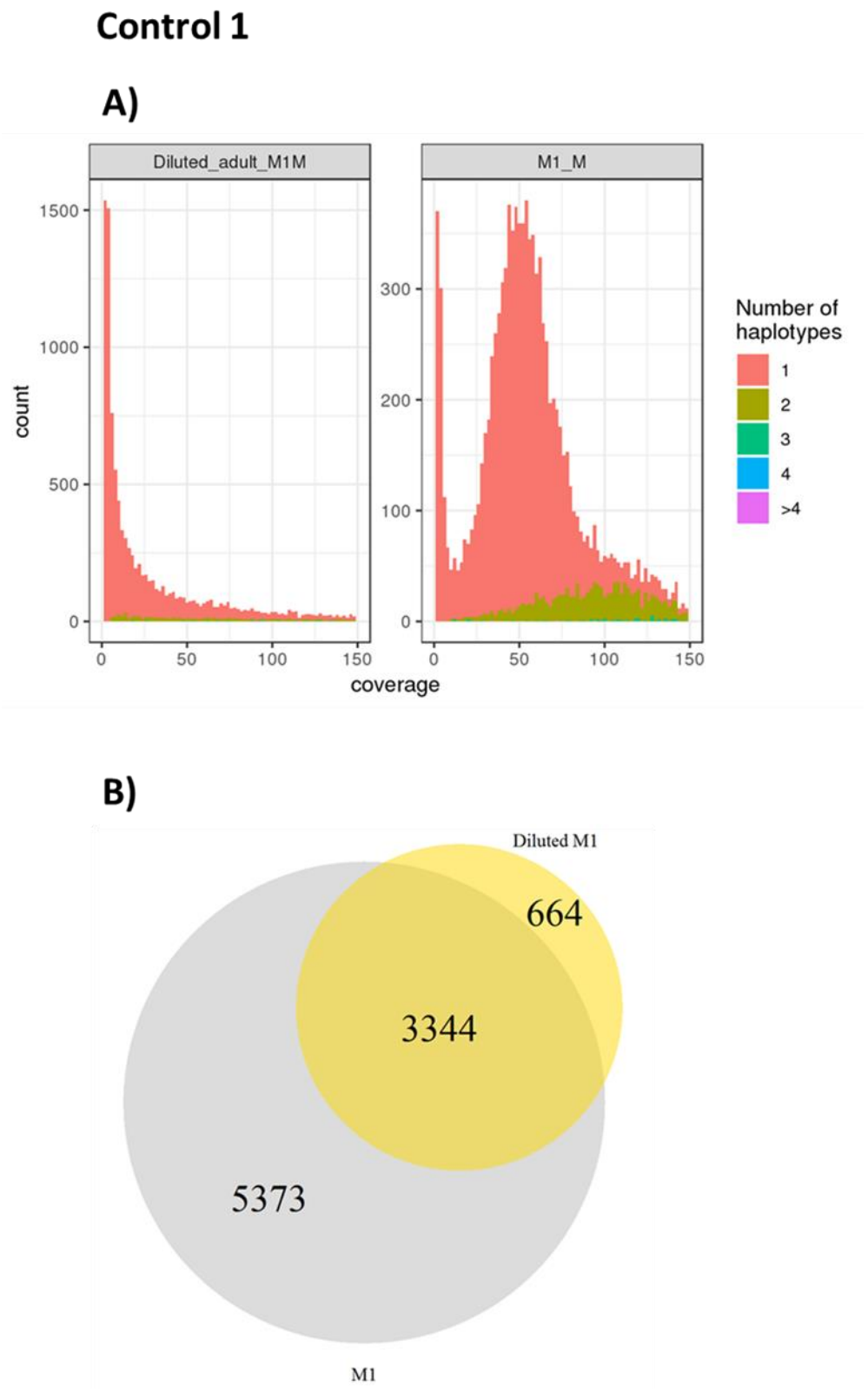


Figure 2. Coverage histogram for the control 1 replicates (A) and the Venn diagram of RAD-loci that are common between the replicate pair (B).

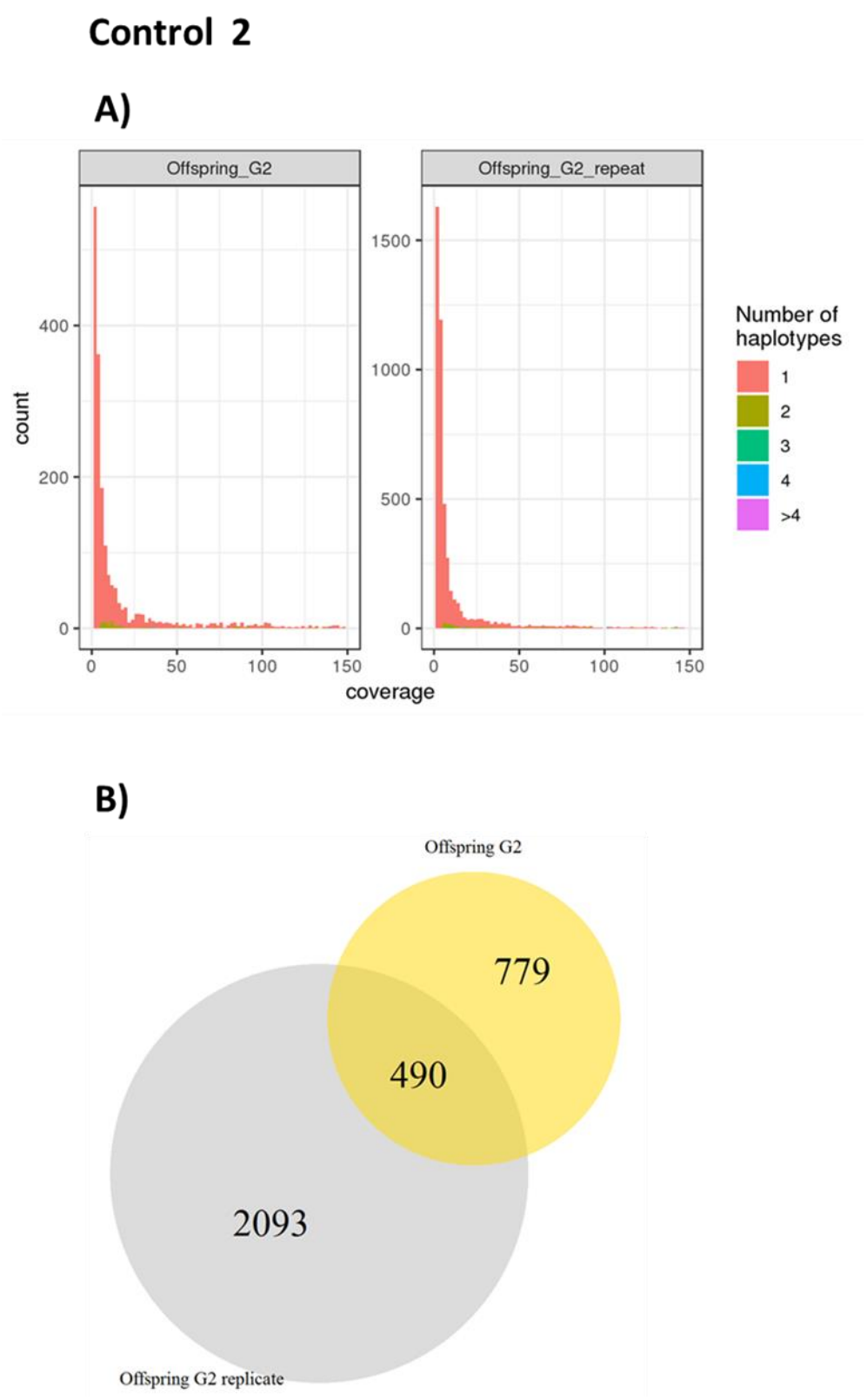


Figure 3. Coverage histogram for the control 2 replicates (A) and the Venn diagram of RAD-loci that are common between the replicate pair (B).

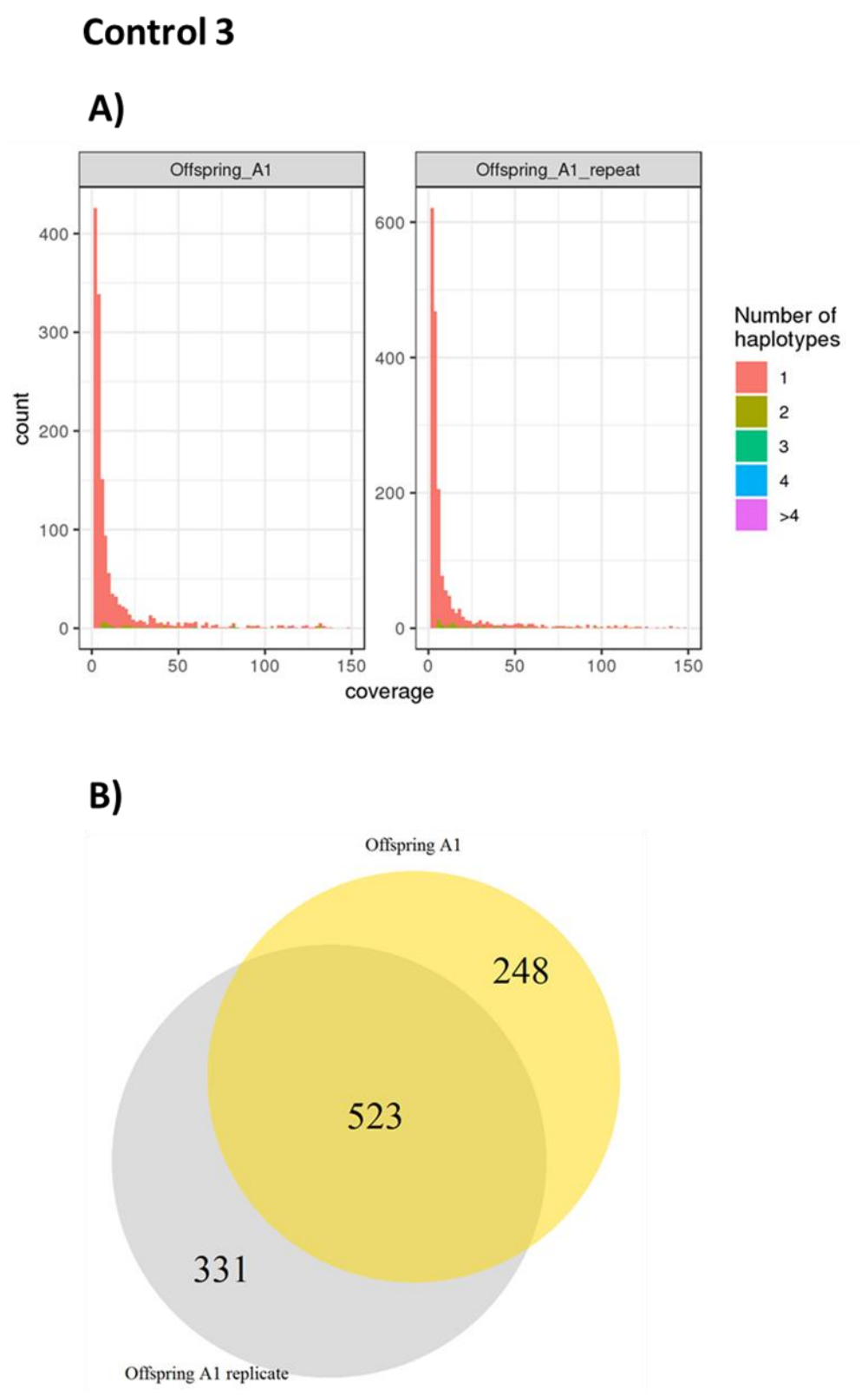


Figure 4. Coverage histogram for the control 3 replicates (A) and the Venn diagram of RAD-loci that are common between the replicate pair (B).

References

CATCHEN, J., HOHENLOHE, P. A., BASSHAM, S., AMORES, A. & CRESKO, W. A. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology,* 22**,** 3124-3140.

CATCHEN, J. M., AMORES, A., HOHENLOHE, P., CRESKO, W. & POSTLETHWAIT, J. H. 2011. Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes|Genomes|Genetics,* 1**,** 171-182.

DEAN, F. B., HOSONO, S., FANG, L., WU, X., FARUQI, A. F., BRAY-WARD, P., SUN, Z., ZONG, Q., DU, Y., DU, J., DRISCOLL, M., SONG, W., KINGSMORE, S. F., EGHOLM, M. & LASKEN, R. S. 2002. Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 5261-6.

HOSONO, S., FARUQI, A. F., DEAN, F. B., DU, Y., SUN, Z., WU, X., DU, J., KINGSMORE, S. F., EGHOLM, M. & LASKEN, R. S. 2003. Unbiased Whole-Genome Amplification Directly From Clinical Samples. *Genome Research,* 13**,** 954-964.

RICHARDS, P. M., LIU, M. M., LOWE, N., DAVEY, J. W., BLAXTER, M. L. & DAVISON, A. 2013. RAD-Seq derived markers flank the shell colour and banding loci of the Cepaea nemoralis supergene. *Molecular Ecology,* 22**,** 3077-3089.